Ph.D. Dissertation of Yoo Duck Kyun

Prediction of antigen binding reactivity using supervised learning on clonal enrichment pattern through bio-panning

지도 학습 기반 바이오패닝 클론 증폭 패턴 분석을 통한 항원 결합 반응성 예측

August 2021

Department of Biomedical Sciences
Seoul National University Graduate School
Biomedical Sciences Major

Yoo Duck Kyun

Prediction of antigen binding reactivity using supervised learning on clonal enrichment pattern through bio-panning

Submitting a Ph.D. Dissertation of
Biomedical Sciences

April 2021

Department of Biomedical Sciences
Seoul National University Graduate School
Biomedical Sciences Major

Yoo Duck Kyun

Confirming the Ph.D. Dissertation written by
Yoo Duck Kyun

July 2021

Chair
Vice Chair
Examiner
Examiner
Examiner

지도 학습 기반 바이오패닝 클론 증폭 패턴 분석을 통한 항원 결합 반응성 예측

지도교수 정 준 호

이 논문을 의학박사 학위논문으로 제출함
2021년 4월

서울대학교 대학원
의과학과 의과학전공
유 덕 균

유덕균의 박사 학위논문을 인준함
2021년 7월

위 원 장 김 종 일

부위원장 정 준 호

위 원 이 민 재

위 원 이 창 한

위 원 심 현 보

Abstract

Background: Monoclonal antibodies (mAbs) are produced by B cells and specifically binds to target antigens. Technical advances in molecular and cellular cloning made it possible to purify recombinant mAbs in a large scale, enhancing the multiple research area and potential for their clinical application. Since the importance of therapeutic mAbs is increasing, mAbs have become the predominant drug classes for various diseases over the past decades. During that time, immense technological advances have made the discovery and development of mAb therapeutics more efficient. Owing to advances in high-throughput methodology in genomic sequencing, phenotype screening, and computational data analysis, it is conceivable to generate the panel of antibodies with annotated characteristics without experiments.

Thesis objective: This thesis aims to develop the next-generation antibody discovery methods utilizing high-throughput antibody repertoire sequencing and bioinformatics analysis. I developed novel methods for construction of *in vitro* display antibody library, and machine learning based antibody discovery.

In chapter 3, I described a new method for generating immunoglobulin (Ig) gene repertoire, which minimizes the amplification bias originated from a large number of primers targeting diverse Ig germline genes. Universal primer-based amplification method was employed in generating Ig gene repertoire then validated by high-throughput antibody repertoire sequencing, in the aspect of clonal diversity and immune repertoire reproducibility. A result of this research work is published in ‘Journal of Immunological Methods (2021). doi: 10.1016/j.jim.2021. 113089’.

In chapter 4, I described a novel machine learning based antibody discovery method. In conventional colony screening approach, it is impossible to identify antigen specific binders having low clonal abundance, or hindered by non-specific phage particles having antigen reactivity on p8 coat protein. To overcome the limitations, I applied the supervised learning algorithm on high-throughput sequencing data annotated with binding property and clonal frequency through bio-panning. NGS analysis was performed to generate large number of antibody sequences annotated with its’ clonal frequency at each selection round of the bio-panning. By using random forest (RF) algorithm, antigen reactive binders were predicted and validated with *in vitro* screening experiment. A result of this research work is published in ‘Experimental & Molecular Medicine (2017). doi:0.1038/emm.2017.22’ and ‘Biomolecule (2020). doi:10.3390/biom10030421’ .

Conclusion: By combining conventional antibody discovery techniques and high-throughput antibody repertoire sequencing, it was able to make advances in multiple attributes of the previous methodology. Multi-cycle amplification with Ig germline gene specific primers showed the high level of repertoire distortion, but could be improved by employing universal primer-based amplification method. RF model generates the large number of antigen reactive antibody sequences having various clonal enrichment pattern. This result offers the new insight in interpreting clonal enrichment process, frequency of antigen specific binder does not increase gradually but depends on the multiple selection rounds. Supervised learning-based method also provides the more diverse antigen specific clonotypes than conventional antibody discovery methods.

Keyword: Antibody discovery, immunoglobulin sequencing, B cell receptor repertoire, high-throughput method, machine learning
Student Number: 2015-22041

Table of Contents

Chapter

List of Figures

List of Tables

1. Introduction

1.1. Antibody and immunoglobulin repertoire

Antibody Structure and Function

B cells play diverse role in immune system of diverse species by recognizing a broad array of antigen via membrane expressed B cell receptors (BCR) and secreted form of the BCR, antibodies[1]. Antibodies are multimeric glycoproteins consists of two copies of heavy chain (about 50 kDa) and light chain (about 25 kDa). These two types of chain share the functionality on highly polymorphic domain, variable region (V_H for heavy chain and V_L for light chain), which recognize the immense number of target antigen. Another functional domain of antibody which having conserved sequence, constant region determines the antibody's architecture and mediates the interactions with effector molecules to elicit unique functionality[2] (Figure 1). After the identification of the antibody structure, important findings have been made in understanding the functions of antibodies in B cell development and humoral immune responses. The genes of V_H and V_L domain are having unique feature which is highly variable in somatic cells, compare to other proteins in genome. Sequence diversity of the variable domain is not equivalently distributed, but focused in specific regions[3]. Both the V_H and V_L domains contain three hypervariable regions neighboring the relatively conserved scaffold, framework regions (Figure 2). All six hypervariable regions, referred to as the complementarity determining regions (CDRs), are forming loop structure and conformationally interact to form the antigen-binding site which is determined by the lowest energy states along with antigen. From the six CDRs, CDR3 has been proven to be the most critical regions which affect the antigen specificity[4]. Previous studies showed that artificially varying other regions while remaining identical CDR3 can maintain the antigen specificity in the variants[5,6].

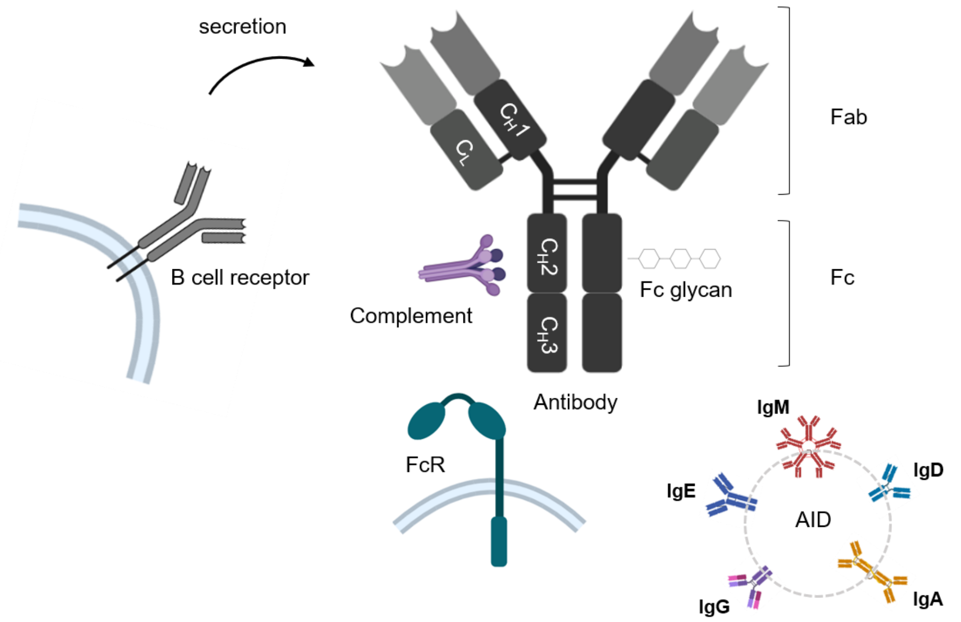

Figure 1. The structure of an antibody and functional domains. The Fab region is responsible of recognizing the various antigen, while the Fc region mediates the effector functions by interact with Fc receptors of the effector cells or activating the complement cascades. Class switch recombination further diversifies the architecture and function of the antibodies.

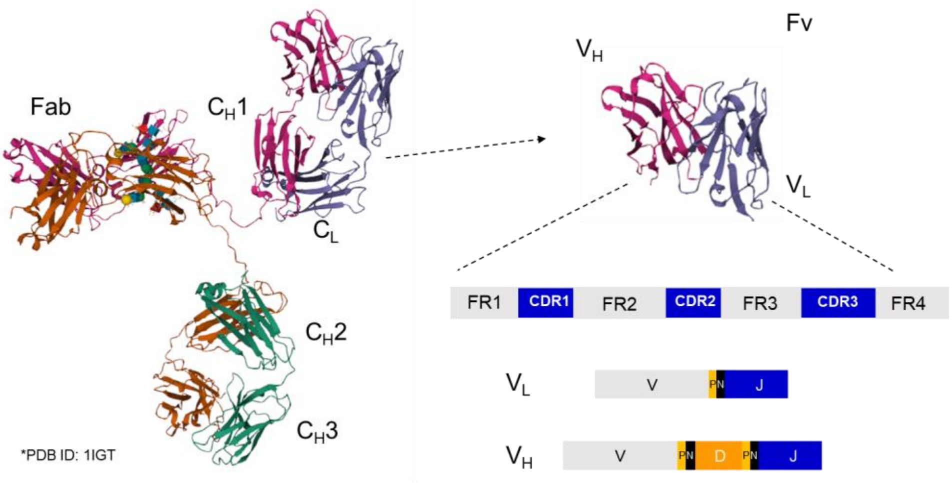

Figure 2. Hypervariable domain of the antibody. The Fab region is subdivided into two domains originated from distinct polypeptide chain of the antibody, variable heavy chain (V_H) and variable light chain (V_L). VH and VL form the fragment of the variability (Fv) which is consisted of sheet shaped region framework regions (FRs) which maintains the scaffold of the antibody, and loop shaped complementary determining regions (CDRs) which responsible of recognizing the various antigen.

Immunoglobulin repertoire

The size of the heavy chain germline locus' is in the order of megabases, and several exons are interspaced by extensive intronic regions. The coding regions of variable domain in the heavy and light chains are composed of multiple gene segments include a variable (V) and joining (J). The heavy chain also contains a diversity (D) segment, surrounded by the V and J gene segments[7] (Figure 3). During early B cell development in the bone marrow, a D and a J gene segment are rearranged to create a continuous exon, and a V gene segment recombines with the DJ complex. There are two light chain loci, kappa (κ) and lambda (λ). The rearrangement of the light chain gene segments occurs after a successful assembly of variable heavy chain gene segments. This recombination process is mainly mediated by the recombinases RAG1 and RAG2[7]. V(D)J recombination is accompanied by insertions and deletions[8]. The insertion process employees several enzymes. RAG1/2 protein generates semirandom P-nucleotides, and terminal deoxynucleotdyl transferase (TdT) adds non-templated N-nucleotides[9]. As a result of the error-prone process, only about 1/3 of rearrangements leading to a functional V(D)J recombination. In the light chain, by the presence of two candidate isotypes, and failure to rearrange one of them on both side of the chromosomes can be remunerated by a successful rearrangement of the other. If all the rearrangement goes unsuccessful, the cell goes through apoptosis progress[7].

After the recombination process in genomic DNA, and upon receiving appropriate activation signals, the successfully recombined V(D)J region splices with the constant (C) gene, in the downstream of the J gene segments. In the heavy chain locus, C gene named mu Cμ, Cδ, Cγ, Cε and Cα are translate into IgD, IgG, IgE and IgA immunoglobulin classes, respectively. The enzyme activation induced cytidine deaminase (AID) triggers the class switch recombination during the germinal center reaction which transduce the activation signal either to the naive and memory B cell. By that, the antigen specificity can be linked with the most effective functions[7]. Along with class switch events, final diversification process is delivered in variable region by antigen dependent manner which take place in germinal center. Activated germinal center B cells, present a high level of transcription, undergo somatic hypermutation (SHM) process which is mediated by AID protein. By SHM, the mutational rate of V regions is raised to 10^{-5}~10^{-3} / base pair compared to 10-9 in the other regions of the genome[10,11]. SHM is based on the DNA repair machinery carries multiple mutations to recover the conversion of cytosine to uridine by deamination[7]. SHM is the mechanism of affinity maturation of the antibodies. Favorable mutations will lead to receptors having improved affinity for the exposed antigen. These B cells carries a survival advantage in clonal expansion and selection to proceed the lateral stage of the B cell development.

AID can engage any nucleotides spanning the whole V(D)J region, but SHM are predominantly detected in the CDRs, which have a major function in antigen recognition[11]. The causes of intrinsic preferential behavior of the AID are not fully elucidated, selection process can result the certain phenomenon[12]. Throughout the entire B cell development, combinatorial and junctional diversity yield a repertoire size of about 5 x 10^{13}, further increases of diversity could be achieved by SHM events. As about 2 x 10^{12} lymphocytes are existed in the human body, the theoretical diversity exceed the individual's diversity capacity[13] (Figure 4).

The last two exons of the constant gene anchor the antibody molecules into the B cell membrane, and translated during the whole B cell differentiation stages. This intermediate cell type is characterized by expression of surface receptors and secreted antibodies simultaneously. and secretion by alternative splicing of a secretory signal domain eliminating the membrane anchoring exons enables the production of shorter form of secreted B cell receptors. Specific cell type plasmablasts secrete antibodies at high levels then terminally differentiate into plasma cells, which lost the surface expression maintaining high level of antibody expression[14] (Figure 5).

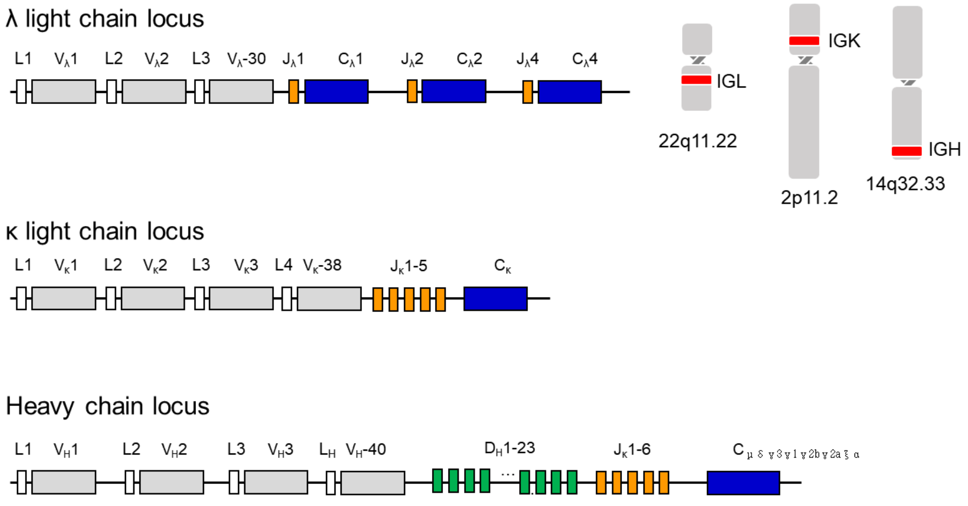

Figure 3. Germline immunoglobulin gene locus. The immunoglobulin gene segments are organized into three genetic loci- kappa (κ), lambda (λ) and heavy (H) chain. Each locus has a multiple gene segments as variable (V), diversity (D) and joining (J) to be recombined to generate genetic diversity during the early B cell development.

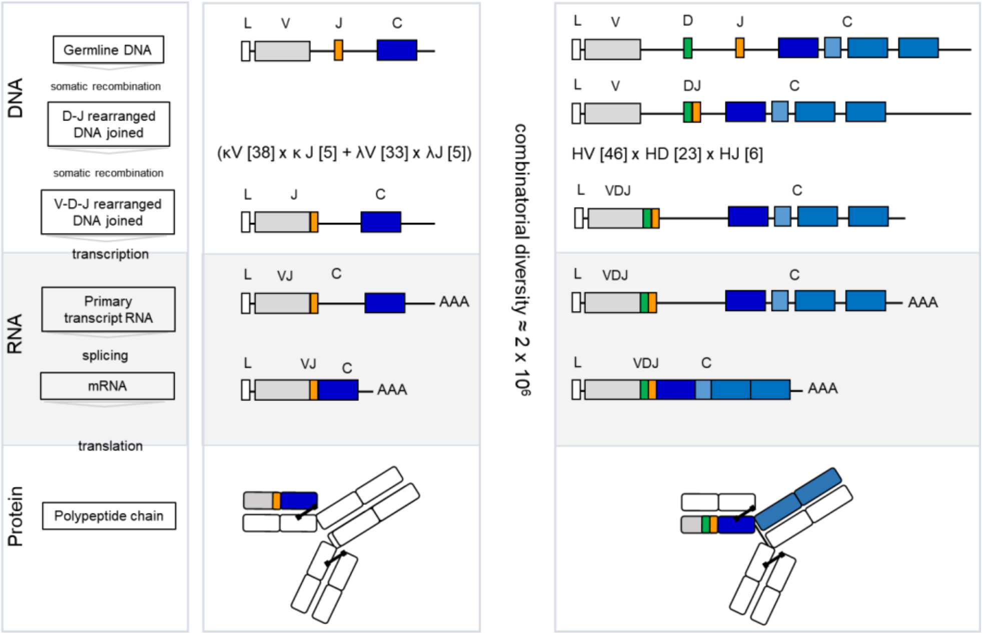

Figure 4. Generation of immunoglobulin repertoire diversity. Recombining gene segments by RAG1/2 protein, and pairing of heavy and light chain, combinatorial diversity is generated (2 x 10^{6}). During the process, additional diversity, junctional diversity is introduced with deletion and addition of the nucleotides by RAG1/2 and TdT enzyme (5 x 10^{13}). After antigen exposure, SHM is introduced which resulting in additional diversity in immunoglobulin repertoire.

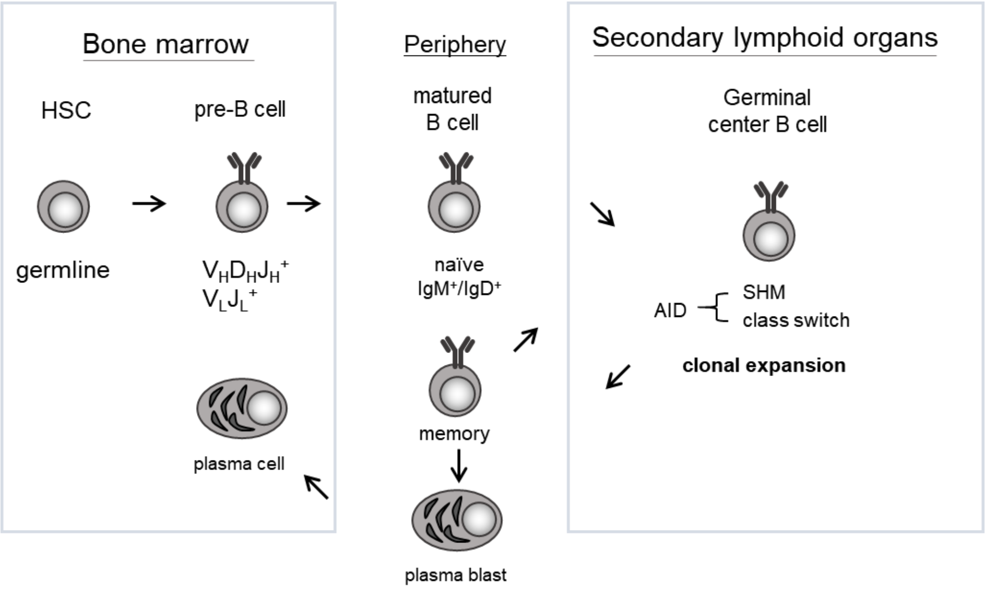

Figure 5. Changes in B cell receptor repertoire following developmental stages of the B cells. Repertoire diversity is generated from the early stage of the B cell maturation by V(D)J recombination along with chain pairing combination in genomic DNA of the pre-B cells. After leaving bone marrow for peripheral maturation of the B cells, naive repertoire is formed to take charge of primary immune responses. After antigen exposure, B cells entered the secondary lymphoid organs including spleen or lymph nodes, facing the further diversification either in variable region and constant region, mediated by AID. Clonally expanded B cells differentiate into memory B cells for secondary immune responses, and plasma cell lineages to elicit systemic humoral responses mediated by secreted antibodies.

1.2. Antibody therapeutics

Monoclonal antibodies (mAbs) and related biotherapeutics

Succeeding the discovery of antibodies and their unique function in human immune system, the first concept of the targeted therapy 'magic bullet' conceived by Paul Ehrlich in the early 1900's[15], became a real-world idea. The history of therapeutic monoclonal antibodies (mAbs) has been developed following hybridoma technology by Köhler and Milstein in 1970's[16]. Hybridoma is generated by fusing non-secretin g myeloma cells with antibody secreting plasma cells[17]. The first monoclonal antibody (mAb) approved for clinical use was Muromonab-CD3 (Orthoclone OKT3) in 1986[18]. The mechanism of action is targeting CD3 co-stimulatory receptor on T cells to desensitize the acute rejection in organ transplantation. Following successful development of therapeutic mAbs, growth of the mAbs market has exploded for the last decades, forecasted to increasing to $125 billion and $ 300 billion in 2020 and 2025 respectively[19]. Globally, about 570 mAbs have been studied in clinical trials for therapeutic usage, and 121 therapeutic mAbs have been marked by the United States Food and Drug Administration (US FDA), 1997-2020[20].

Over the native immunoglobulin molecule as mAb therapeutics, related therapeutic product field continues to expand as antibody drug conjugates (ADCs) and antibody derivative molecules (Figure 6). By conjugating cytotoxic drug or enzymatic payload to mAbs, it is conceivable to achieve therapeutic efficacy and target specificity at the same time[21]. Recombinant bi-specific antibodies can engage distinct targets to induce synergistic effect of multiple mechanisms of actions. Also, physical linking of target cells or molecules is possible using bi-specific antibodies. The bispecific T cell engager (BiTE), can target pathogenic cells such as tumor, by bringing T cells to efficiently localize the cytotoxic effect[22]. Other breakthrough application of mAb combined with cellular therapy, chimeric antigen receptor (CAR)-T cell therapy became an advancement in cancer treatment[23]. CAR-T is generated by genetical engineering of T cells to express a synthetic receptor such as binding domain of the mAb followed by cellular expansion to be infused into the patient's body to attack pathologic targets. CAR-T targeting CD19 target showed dramatic clinical response and high rates of complete remission have been observed in the setting of B cell malignancies, acute lymphoblastic leukemia and diffuse large B-cell lymphoma, resulting in four FDA approvals[24]. In 2020, 191 active pre-clinical and clinical trials were directed at CD19 to improve the efficacy and safety[25]. Also, other challenges are ongoing to develop CAR-T cell therapy to address “off-the-shelf” allogeneic therapy, engineering strategies and aiming next-generation targets, BCMA, CD20, CD22 and HER2 to overcome the limitations and resistances[26-29].

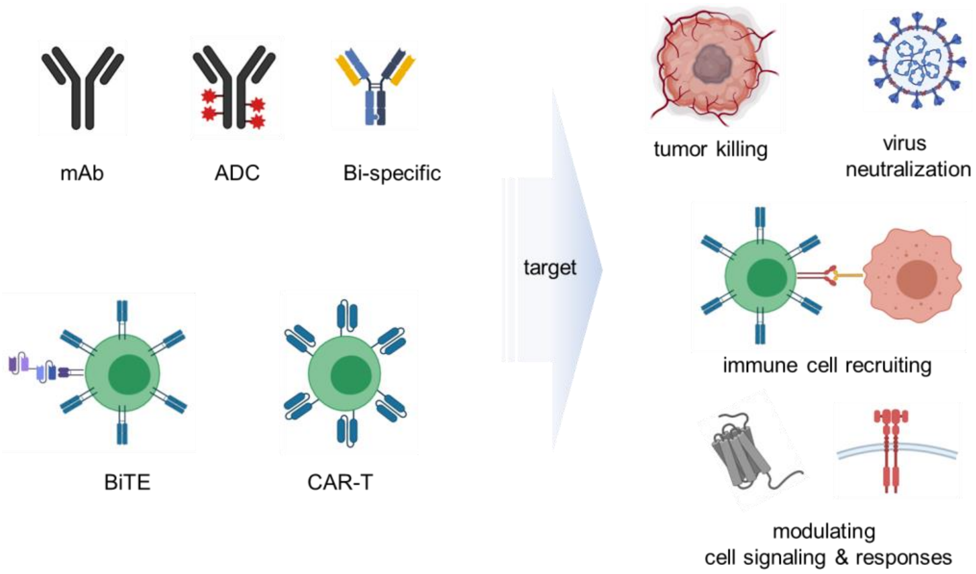

Figure 6. Monoclonal antibody and related biotherapeutics. By targeting various systems in pathogenesis, it is conceivable to conduct targeted therapy using mAb and related biotherapeutics. By combining other drug classes, antibody-drug conjugates (ADC), which carries traditional small molecule drugs as a payload in the antibody scaffold, and chimeric antigen receptor T cell (CAR-T) and bispecific T cell engager (BiTE) platform, have been showing remarkable clinical responses compare to native form of the mAbs.

Developmental process of antibody therapeutics

Since the mAb market value has been rapidly growing and proving the therapeutic potential in multiple disease area, myriad number of academic and industrial competitors started to discover therapeutic targets and establish advanced technologies in antibody drug development. The conventional development cycle of an antibody discovery campaign can be subdivided into distinct steps. 1) candidate discovery, 2) lead optimization, 3) pre-clinical development, and 4) clinical trials[30]. (Figure 7) After identification of target-specific antibodies, initial candidates then enter the series of engineering procedure to improve their antigen binding properties (affinity, specificity, cross-species reactivity) and druggabilities (physicochemical properties which is translated into thermal/chemical stability, aggregation propensity, productivity, solubility, solubility, and immunogenicity)[31]. Antibody discovery and engineering stage is of critical importance for a poor metric of sequence derived biophysical property can lead to failure of downstream development process[19,32].

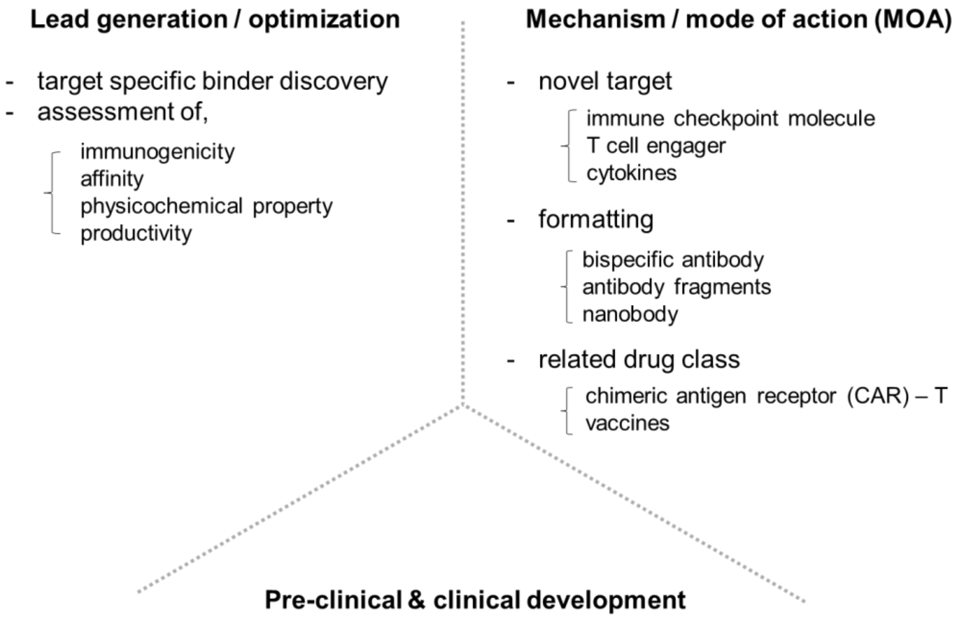

Figure 7. General contemplational process in development of antibody therapeutics. The first step in antibody therapeutics development is generation of target specific binders and meta data of the antibody characterization. In the parallel procedure, MOA study is essential for the further clinical development stages.

1.3. Methodology: antibody discovery and engineering

Conventional antibody discovery and engineering methods

Back in 1975, Georges J. F. Köhler and César Milstein established a method to generate hybridoma cells by fusion of immortalized B cells and myeloma cells. Hybridoma cells originated from an immunized organism could be used in selection of antibody secreting properties against the specific antigen used in immunization. After selection, antigen specific clones could be entered the scale-up process to generate large amount of antibody molecules[33]. Although, the hybridoma technology has technical limitations such as laborious, low efficient limiting dilution subcloning process and intrinsic limitation, immunogenicity, which could be unsafe in therapeutic usage. To overcome the certain limitations of the hybridoma technology, in vitro based antibody discovery methods have been developed.

By harvesting genomic materials from the living organ having adaptive immune system, it is possible to reconstitute the antibody repertoire in vitro by cloning immunoglobulin genes into certain antibody display systems. The concept of antibody display is consisted of several key elements. (1) Construction of antibody library carrying the diversity at the genotypic level, (2) linking genotype to phenotype by utilizing in vitro display system and (3) applying selective pressure to screen out antigen specific binders[34]. Various type of in vitro display libraries can be generated from a variety of hosts for adequate purposes. Naive libraries are not biased towards the specific antigen stimulation so that carries higher diversity in repertoire. In contrast, immune libraries from an immunized (or infected) are less diverse, but having a significant level of enrichment in antigen specific binders. To complement the limitations of each libraries, synthetic and semisynthetic libraries are developed for improvement of the library features[35].

The first in vitro display system is translated by phage play platform (Figure 8). The finding that recombinant peptides and proteins can be expressed upon fusion with the coat proteins resulting in display of the intended molecules on the surface of filamentous bacteriophages[36]. With the antibody displayed on the surface of the phage, the displayed antibodies can bind to an immobilized antigen enabling the selection of antigen specific clones. After washing out non-binding phage particles, antigen specific phage particles are eluted identified by phagemid sequencing. The major limitation of phage display platform is that the identified antibody clones enter the mammalian expression system which carries post-translational modification which could affect the expression level or even in antigen binding properties[37]. Despite the certain limitations, due to its' robustness and effectiveness, 9 mAbs are approved on the market and more than 20 phage-display derived mAbs are in late-stage clinical trials[19,38].

Eric Boder and Dane Wittrup suggested the yeast display platform to overcome the limitations of the phage display system in 1997[39]. Yeast (*Saccharomyces cerevisiae*), as a eukaryotic organism, present the post-translational modifications, protein folding, and secretory machinery similar to mammalian system. By cloning of immunoglobulin genes into a yeast expression plasmid and transformed to generate antibody libraries, slightly smaller than phage, of 10^7 to 10^9 diversity[40], it is conceivable to fusing antibodies to the Aga2p protein, similar to phage display system.

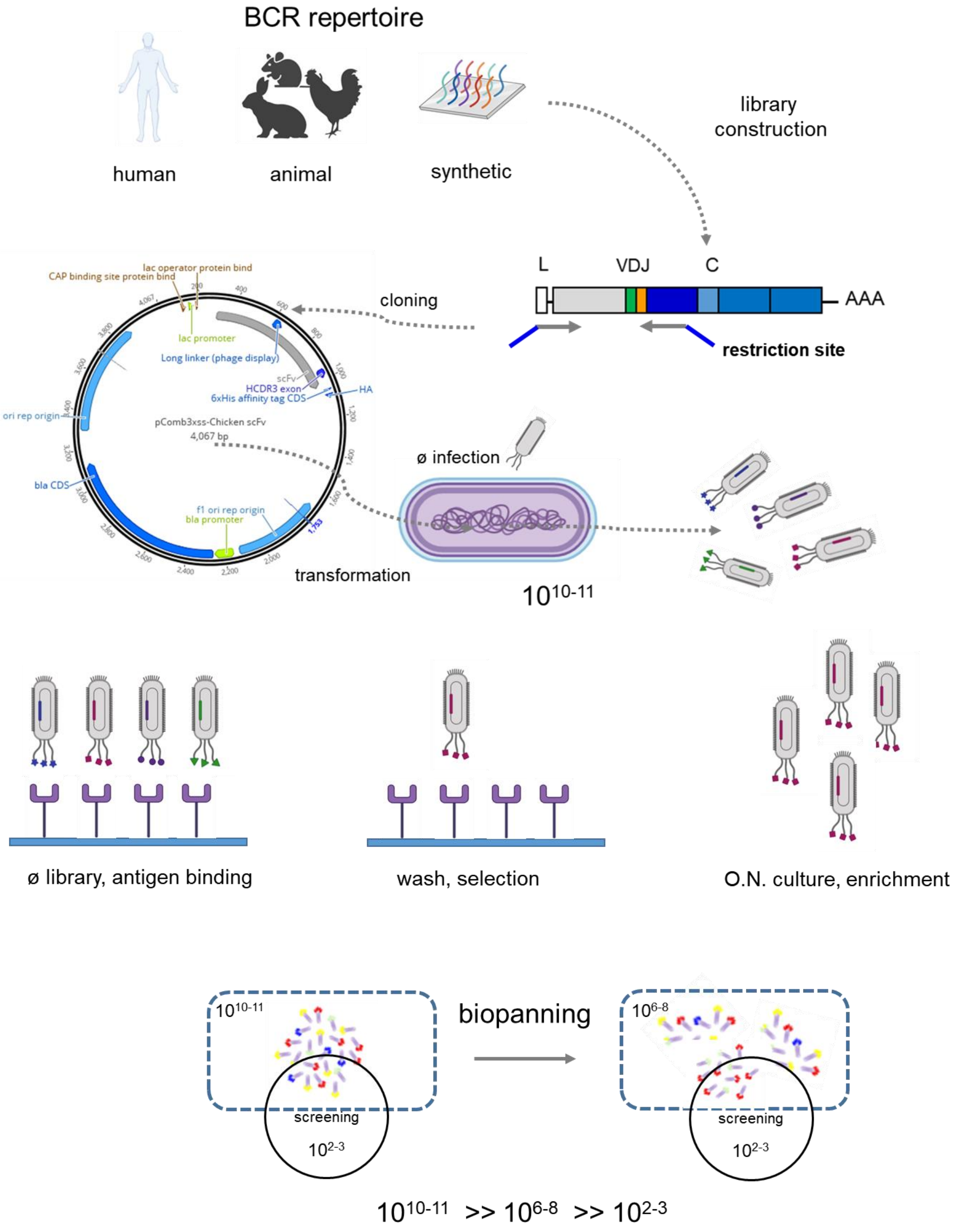

Figure 8. Phage display and biopanning. Recombinant DNA technologies enables the physical translation of genetic antibody repertoire to surface displayed antibody protein repertoire which could be subjected into selective cycles against specific antigen. Repetitive enrichment process of the antigen specific clones, biopanning, makes the screening of binders from huge sized antibody repertoire.

Transgenic humanized animal platform

While the hybridoma, in vitro display technologies represented a certain advance in development of antibody therapeutics, the candidate originated from non-human organism eventually led to high immunogenicity. To overcome the limitations, generation of chimeric and humanized sequence of antibodies are developed[41]. Chimeric antibodies are consisted of a non-human Fab with a human Fc region and humanized antibodies are generated by grafting CDRs of the antigen specific non-human antibodies to human framework scaffold. which graft murine[41]. Bur during the process, binding properties of the original clones could be modified resulting in further engineering of binding affinity which is lagging the developmental process.

To improve the technologies generating fully human antibodies, the demonstrations that human immunoglobulin gene loci could be introduced into the mouse genome[42]. These efforts resulted in the first fully human transgenic mouse model, XenoMouse® and HuMab-Mouse® and have been followed by a series of, humanized animals[43-45]. The groundbreaking ability of these platform to understand non-human immune system and human antibody repertoire has revolutionized biotechnology by providing a diverse source of fully human antibodies. Also, fully humanized antibodies derived from these platforms are taking a place in clinical usage and development[46].

Third generation method: deep mining of antibody repertoire by next-generation sequencing

With tremendous advances in antibody discovery and engineering platforms are emerged, screening of antibodies under high-throughput manner is the most key and fundamental parts in antibody drug development. Sanger sequencing is commonly used at each of technologies reviewed in the previous section which could limit the scale and throughput of the antibody discovery platform. Even the phage display platform which leverages the most diverse antibody repertoire in selective procedure, isolation of individual phagemids from pools is limited to sequencing a few hundred to thousand clones generally. Since phage display gives an output of up to ~10^8 sequences, the use of high-throughput sequencing (HTS) or next-generation sequencing (NGS) becomes essential to interrogate the entire sequence diversity. Recent advances in high-throughput DNA sequencing technologies over the last decade has led to a dramatic reduction in the cost of sequencing and has revolutionized the development of the computational methods in immunobiology[47]. By combining bioinformatic tools analyzing clonal selection procedure, it is conceivable to identify the antigen specific antibodies without experimental procedures[48].

Not only in analyzing in vitro display repertoire and selection procedure, the NGS analysis has enabled comprehensive characterization of the B cell receptor (BCR) landscape at various aspects[49] (Figure 9). Large-scale computational structural modeling has revealed the correlation of the sequence and structural between naive and antigen-experienced antibody repertoires[50]. NGS-aided profiling of the BCR repertoire of multiple diseases such as infectious disease, cancer and autoimmune disease can provide the comprehensive understanding in antibody mediated pathogenesis, which could be directly applied to develop the therapeutics and diagnostics.

Recently, by analyzing shaping of the B cell response in HIV-infected individuals, Jardine et al. has developed the engineered immunogen that could target B cells that express the germline clonotypes originated from a particular broadly neutralizing antibodies (bNAbs) class[51]. In the case of coronavirus disease-19 (COVID-19), Kim et al. have identified the stereotypic neutralizing antibody against SARS-CoV-2 in healthy individuals and COVID-19 patients. They isolated SARS-CoV-2 spike protein receptor binding domain (RBD)-specific clonotypes composed of immunoglobulin heavy variable 3-53 (IGHV3-53) or IGHV3-66 and immunoglobulin heavy joining 6 (IGHJ6) genes in COVID-19 patients. These clonotypes were also detected in more than half of the heathy cohorts, which provide the evidence of the pre-existing neutralizing antibodies in naive BCR repertoire[52]. By analyzing BCR repertoire of the cancer patients who received the immune checkpoint inhibitor (ICI)and showed the clinical response, highly convergent antibody repertoire was detected and shared between the multiple patients[53]. A recombinant antibody which reconstituted from the convergent BCR sequences were treated with ICI drug as a combination therapy, showed improved therapeutic efficacy compared to the single treatment of the ICI drug[54]. Further, disease types that are not deeply studied in the related mechanism of the adaptive immunity, including neurodegenerative disease, became emerging targets in BCR repertoire profiling studies.

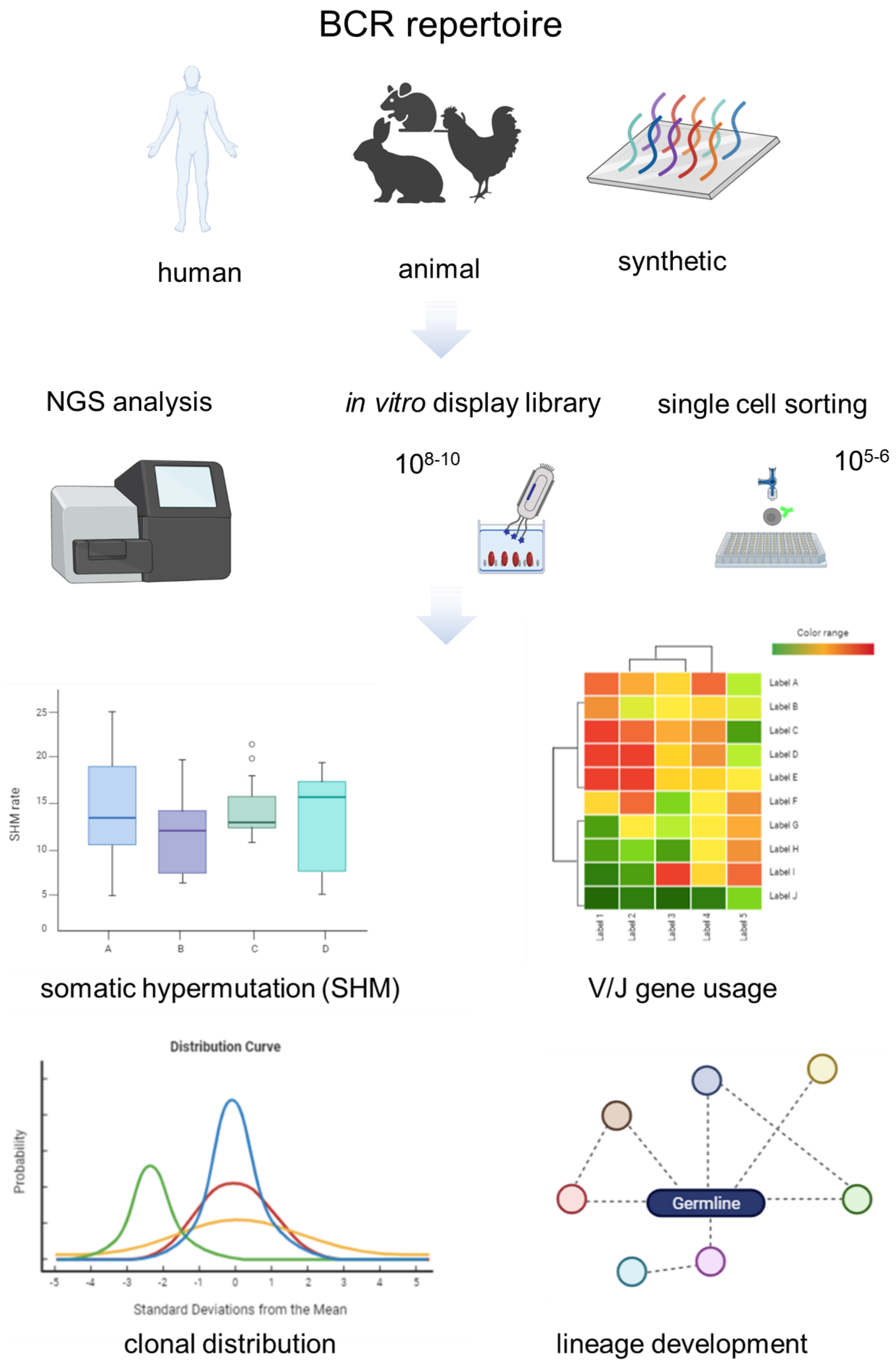

Figure 9. B cell receptor repertoire analysis with next-generation sequencing. Experimental and computational improvements in BCR repertoire profiling enables the next-generation antibody discovery.

2. Thesis objective

This thesis aims to develop the novel next-generation antibody discovery method by combining phage display, immunoglobulin sequencing and high-throughput clone retrieval method.

Improving conventional phage display library construction method

To construct antibody libraries from B cells, singleplex or multiplex PCR amplification were conducted using primers targeting multiple immunoglobulin genes. However, during this process, the B cell receptor (BCR) repertoire is distorted due to interactions between multiple target genes and primers. To overcome the conventional limitations, I devised new way of library construction method which minimize the Ig gene amplification bias.

Identification of antigen specific antibodies using immunoglobulin sequencing data derived from biopanning, and high-throughput clone retrieval method

In conventional phage display and biopanning methods, it is known that the critical drawback of the method is extremely low binder screening efficiency. Utilizing phage enzyme-linked immunoassay (ELISA) and Fluorescence-activated Cell Sorting FACS resulting in 10^{2-3} screening scales from 10^{9-11} diversity of the initial repertoire. To fully take advantage from hypervariable antibody repertoire, we generate *in silico* sequence data annotated with binding property. Supervised machine learning was employed to annotation of the binding property along with experimental screening data obtained from high-throughput clone retrieval method.

3. Establishment of minimally biased phage display library construction method for antibody discovery

This is an author produced adaptation of a peer reviewed article published in

Journal of Immunological Methods (2021). doi: 10.1016/j.jim.2021. 113089.

[Lee Y*, Yoo DK*, Noh J*, Ju S, Lee E, Lee H, Kwon S & Chung J]

* These authors contributed equally.

3.1. Abstract

Immune hosts are valuable sources for antibody discovery. To construct antibody libraries from B cells, singleplex or multiplex PCR amplification were conducted using primers targeting multiple immunoglobulin genes. However, during this process, the B cell receptor (BCR) repertoire is distorted due to interactions between multiple target genes and primers. To minimize this alternation, we devised a new method for harvesting immunoglobulin genes and tested its performance in rabbit VH and VK genes. Double-stranded cDNA was synthesized using primers containing V/J gene-specific and universal sequence parts from B-cell RNA. VH and VK gene libraries were obtained through subsequent PCR amplification using primers with universal sequences. Next-generation sequencing analysis confirmed that universal primer PCR libraries had more diverse VH and VK gene repertoires, more clonotypes retrieved from the BCR repertoire of RNA samples, and a higher relative frequency correlation than conventional singleplex or multiplex gene-specific primer libraries.

3.2. Introduction

Immunized or infected organisms are great sources for antibody discovery as the antigen-dependent clonal selection and expansion of B cells occurs *in vivo* [1]. Antibody display libraries have been constructed from a wide variety of species possessing a humoral immune system including humans [55–58], mice [59], rabbits [60,61], cow [62], chickens [63,64], and sharks [65]. To take full advantage of the indigenous *in vivo* system, it is essential to construct an antibody library that accurately reproduces the host B cell receptor (BCR) repertoire.

Due to the rapid advances of next-generation sequencing (NGS) technologies, the germline variants of immunoglobulin (Ig) genes have been identified in multiple species, leading to the successful design of gene amplification primers [66]. The typical structure of Ig genes contains multiple variable (V), diversity, and joining (J) gene segments for heavy chains or V and J gene segments for light chains. Therefore, multiple primer sets have been designed to cover the entire BCR repertoire in the construction of an antibody library.

Conventionally, Ig genes can be amplified using multiple primers with either singleplex or multiplex amplification. In singleplex amplification, every set of forward and reverse primers is employed for each individual polymerase chain reaction (PCR). Then, the amplicons from each PCR are pooled. In multiplex amplification, library construction is conducted in a single PCR reaction using a mixture of all the forward and reverse primers, which enables the handling of multiple samples in a more high-throughput manner. In addition, multiplex amplification is unavoidable when the quantity of genomic materials is limited. However, cross-priming of primers to unintended immunoglobulin genes during multiplex amplification can skew the clonal distribution [67,68]. Furthermore, multiplex amplification has more challenges such as cross oligonucleotide dimerization, different primer annealing temperatures, and preferential amplification of specific targets caused by steric interference among multiple primers [69–71]. Despite these explicit limitations in precisely replicating the BCR repertoire, antibody display libraries have been constructed using Ig gene sequence-specific primers [56-58,72-79]. However, the extent of the bias introduced in these libraries has not been studied in detail.

To reduce the bias introduced during the PCR amplification of Ig genes using sequence-specific primers, we devised and tested a new way to harvest Ig genes. First, double-stranded cDNA was prepared using primers containing the Ig gene–specific sequence and universal sequence parts from RNA samples of rabbit B cells. Then, heavy chain variable region (V_H) and kappa light chain variable region (V_K) genes were amplified using universal primers through PCR. NGS analysis confirmed that the V_H and V_K gene repertoires are more diverse in these libraries. In addition, a higher number of clonotypes were retrieved from RNA samples with a higher correlation of relative frequency than conventional singleplex or multiplex libraries using gene-specific primers (Figure 10).

3.3. Results

Construction of antibody libraries

We devised a new way of harvesting V_H and V_K genes (Figure 1). First, RNA was isolated from spleen, bone marrow and peripheral blood mononuclear cells of a rabbit and subjected to reverse transcription reactions with primers containing (1) a J gene-specific sequence, (2) a *Sfi*I restriction site or linker, (3) a unique molecular identifier (UMI) barcode for the precise error correction [80], and (4) a partial reverse Illumina adaptor (P7). After the reverse transcription reaction, the second strand cDNA was synthesized with multiple forward primers encoding (1) an Illumina adaptor (P5), (2) a V gene-specific region, and (3) a *Sfi*I restriction site or linker. Then, the double-stranded cDNA was subjected to PCR amplification with universal primers targeting the conserved regions of the templates (P5, P7). Thereafter, we confirmed the successful construction of the single chain variable-fragment (scFv) gene library through overlap extension PCR using these amplified V_H and V_K genes (Figure 11).

In parallel experiments, conventional V_H and V_K libraries were prepared from the same RNA sample through both singleplex and multiplex amplification using gene-specific primer sets, which were rationally designed from the international ImMunoGeneTics information system (IMGT) database by Peng et al [81]. As a result, three V_H libraries (singleplex PCR using universal primers, singleplex PCR using gene-specific primers, and multiplex PCR using gene-specific primers) and three V_K libraries were prepared then subjected to NGS analysis for comparative evaluation using the reference data. The reference data which accurately reproduce the V_H and V_K repertoire in the RNA sample were generated under UMI-based error correction (Table 1), which has been widely used in BCR repertoire analysis [82-84].

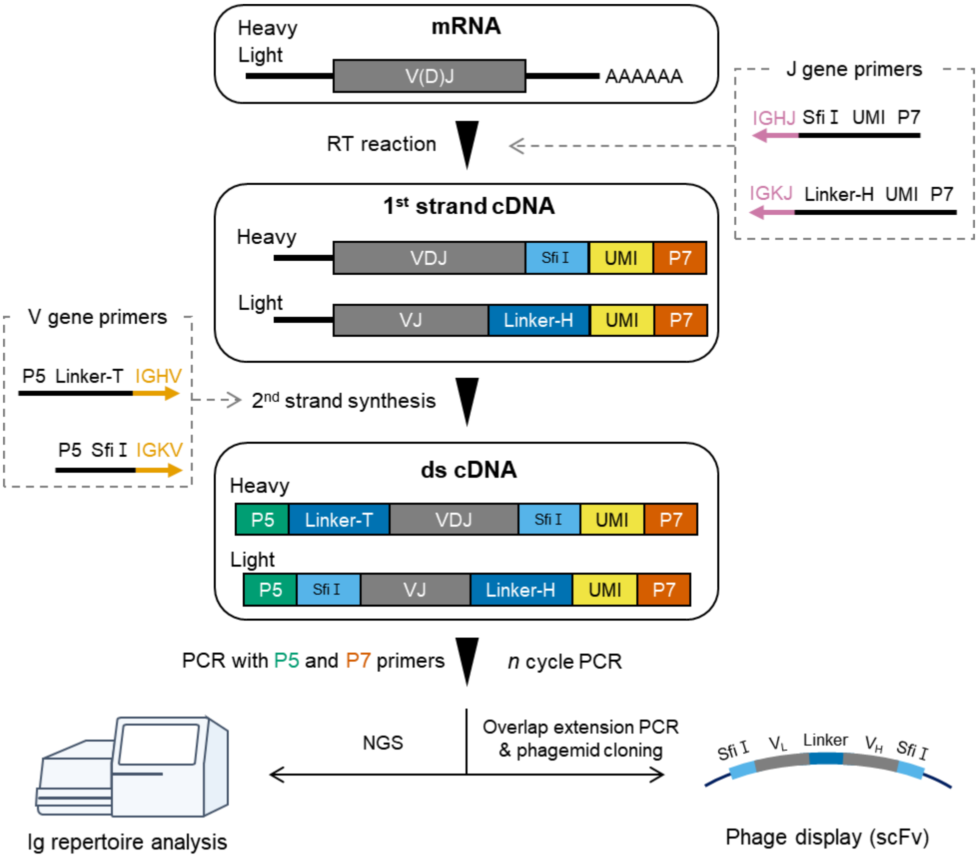

Figure 10. Flow chart for the construction of the antibody display library. The process combined two steps, double-stranded cDNA synthesis and PCR using universal primers. The PCR products were subjected to either next-generation sequencing (NGS) analysis or overlap extension PCR. The partial linker regions, linker-head (H) and linker-tail (T), are used for linking V_K-V_H amplicons during overlap extension PCR. ds, double strand; Ig, immunoglobin; RT, reverse transcription; scFv, single-chain variable fragment; UMI, unique molecular identifier.

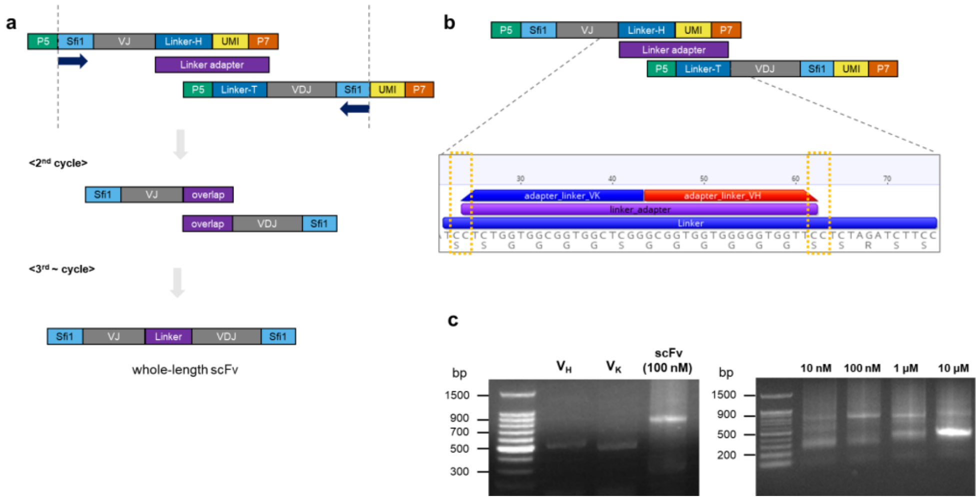

Figure 11. Construction of scFv library via overlap extension PCR. (a) Construction of scFv gene through overlap extension PCR amplification. The sequence of the primers used in overlap extension PCR are listed in supplementary table 3. (b) Linker adapter primer was designed to avoid intramolecular annealing (considered sequence regions are highlighted as dotted yellow boxes) while sharing similar melting temperature of amplification primers targeting Sfi I restriction site. (c) A representative 1% agarose gel electrophoresis showing scFv product after overlap extension PCR for scFv (left panel). Concentration of the linker adapter primer used in PCR reaction was optimized (right panel).

Table 1. Statistics for the preprocessing results of the NGS data. In this research, 8 samples were prepared and NGS data were constructed. Heavy chain and kappa chain of rabbit B cell receptor (BCR) transcript were amplified with different methodologies; singleplex, multiplex, and the alternative method that we propose in the article. Other libraries of heavy and kappa chain were prepared, for the construction of reference libraries which were error-corrected with UMI processing.

Sample name	Raw reads	Functional reads	Unique functional reads (clones)
Heavy_singleplex	3,269,549	1,430,067	170,172
Heavy_multiplex	3,634,768	1,483,354	164,323
Heavy_alternative	2,055,213	446,602	109,047
Heavy_reference	2,394,909	75,740	56,483
Kappa_singleplex	3,354,238	769,814	100,644
Kappa_multiplex	3,411,854	959,708	116,297
Kappa_alternative	3,028,775	1,090,895	150,884
Kappa_reference	2,213,801	68,060	40,885

Universal PCR primer libraries achieved higher diversity

We determine the diversity of V_H and V_K libraries by Hill numbers, which have been widely used in BCR repertoire analysis [85-89]. Hill numbers (qD) are the integration of diversity indices differing by an exponent q (q ≥ 0), defined as

$$^{q}D = \left(\sum_{i=1}^{S} p_i{}^{q}\right)^{\frac{1}{(1-q)}}$$

where S is the whole number of unique Ig clonotype species and pi is the frequency of regarding species. We used Hill numbers of q = 0, 1, and 2, which represent species richness, Shannon diversity, and Simpson diversity, respectively, in the quantification of the diversity of the libraries [90].

Species richness is the number of different sequence components without considering their frequencies. Because species richness is highly sensitive to sample size [90], balancing NGS throughput among libraries is a prerequisite for a fair comparison. To normalize the NGS throughput of the libraries, the sample coverage of individual NGS data was calculated (Figure 12a). Sample coverage is defined as the proportion of Ig sequences covered by a specific NGS throughput. In addition, interpolate and extrapolate estimates for the sample coverage can be calculated as varying NGS throughput. We selected a NGS throughput where the sample coverage of the libraries equals each other: 83.8% sample coverage in V_H, and 91.5 % sample coverage in V_K (Figure 12a). Then, the unique sequence component was defined at the clonotype level as sequences containing the same V and J gene and showing identical amino acid sequences at the complementary determining region 1, 2, and 3 (CDR1, 2, and 3). At the level of clonotype, species richness values were higher in universal PCR primer libraries (V_H, 1.62-fold; V_K, 2.27-fold) than singleplex and multiplex gene-specific PCR primer libraries (V_H, 1.58-fold; V_K, 2.06-fold) (Figure 12b).

Shannon diversity and Simpson diversity weight the frequency of each sequence component and can be interpreted as the number of typical clonotypes with moderate clonotype frequency and the number of dominant clonotypes with high clonotype frequency, respectively. Shannon diversity values of universal PCR libraries were similar to those of singleplex and multiplex gene-specific PCR primer libraries: 1.15-fold and 1.08-fold for heavy chain, and 0.99-fold and 1.07-fold for kappa chain, respectively (Figure 12c). Simpson diversity values of universal PCR primer libraries were much lower: 1.74-fold and 1.70-fold for heavy chain, and 1.46-fold and 1.49-fold for kappa chain, respectively (Figure 12d). The comparison was conducted with the asymptotic values of each diversity estimate.

The higher values of species richness but lower values in Shannon and Simpson diversity achieved in the universal PCR primer libraries indicated that rare clonotypes were more effectively retrieved from the RNA sample and typical and dominant clonotypes were not over-represented in comparison to gene-specific PCR primer libraries.

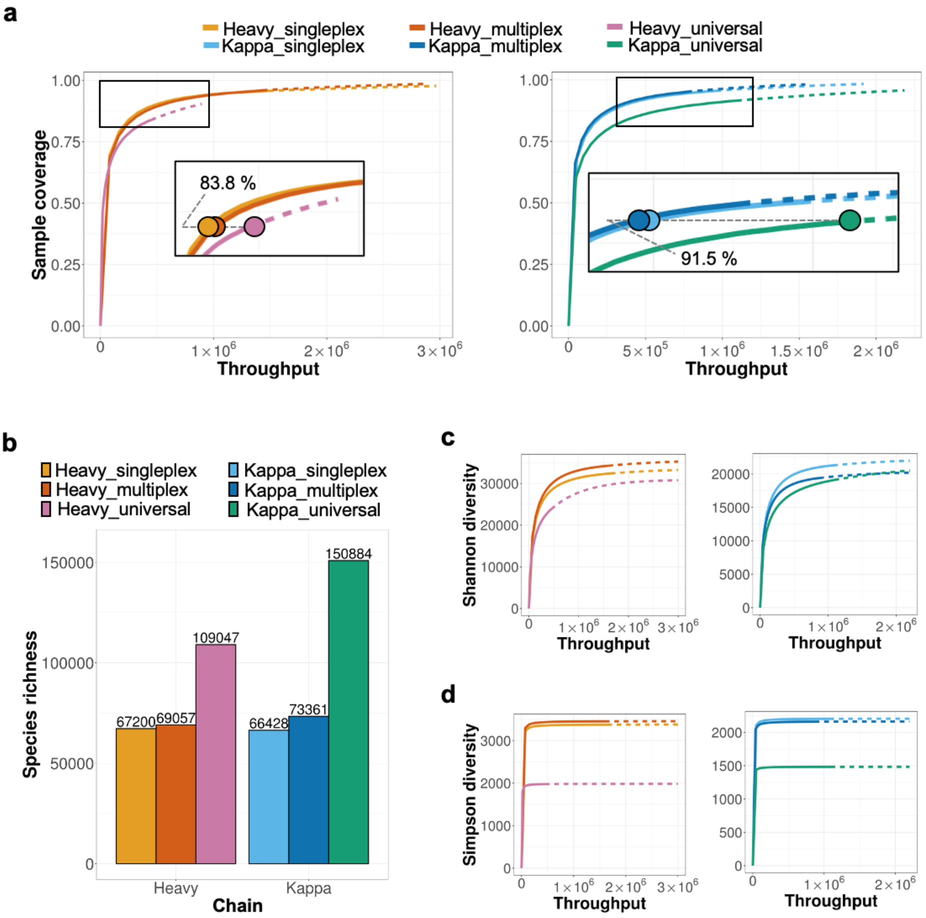

Figure 12. Clonal diversity among libraries. (a) Sample coverage estimates of the libraries and the throughput balancing. Sample coverage estimates were calculated using iNEXT R package, and the throughput to be sampled was chosen by the point at which three libraries had the same sample coverage estimate. (b) Species richness from V_H and V_K libraries amplified by each method. The values were calculated from throughput-balanced data. (c-d) Shannon diversity estimates and Simpson diversity estimates of the libraries. In Figure 2a, 2c, and 2d, the solid line represents the rarefaction curve, and the dotted line represents the extrapolation curve.

Universal PCR primer libraries retrieved a higher number of clonotypes with a higher relative frequency correlation from the RNA sample

To determine the efficiency in retrieving V_H and V_K clonotypes among Ig libraries, we compared their similarity in V/J gene usage to the reference data on the Ig repertoire from an RNA sample, counted the number of overlapping clonotypes, and analyzed the correlation in the frequency of overlapping clonotypes. The V/J gene usage of each library was computed for V_H (Figure 13a) and V_K (Figure 13b). To quantify their similarity to the reference data, cosine similarity was calculated to measure the angle between two multidimensional (39 heavy variable, 6 heavy junctional, 24 kappa variable, and 5 kappa junctional) vectors ranging from −1 (exactly opposite) to 1 (exactly the same). The results showed that the cosine similarity of all libraries exceeded 0.9, implying that all libraries effectively reproduced the original V/J gene usage proportion.

The number of overlapping clonotypes between the reference data and each individual library was calculated (Table 2). Only 14.5% V_H and 28.4% V_K clonotypes overlapped with those in the reference data for the singleplex gene-specific PCR primer library. For the multiplex library, the results were similar (15.0% V_H and 28.5% V_K). For the universal PCR primer libraries, the overlapping percentage significantly increased to 29.6% for V_H and 56.8% for V_K. In particular, for the top 1,000 clonotypes in the RNA sample selected on the basis of their frequency, more than 70% existed in the top 1,000 clonotypes of universal PCR primer libraries. However, the percentage dropped to less than 33% and 51% in gene-specific PCR primer V_H and V_K libraries, respectively.

Afterward, the correlation in frequency among overlapping clonotypes was checked (Figure 13c). The R^2 values were higher than 0.95 in universal PCR primer libraries. However, R^2 values were lower than 0.56 and 0.82 in gene-specific PCR primer V_H and V_K libraries, respectively. These results revealed that universal PCR primer V_H and V_K libraries were significantly superior to gene-specific PCR primer libraries in both retrieving more clonotypes and maintaining their relative frequency.

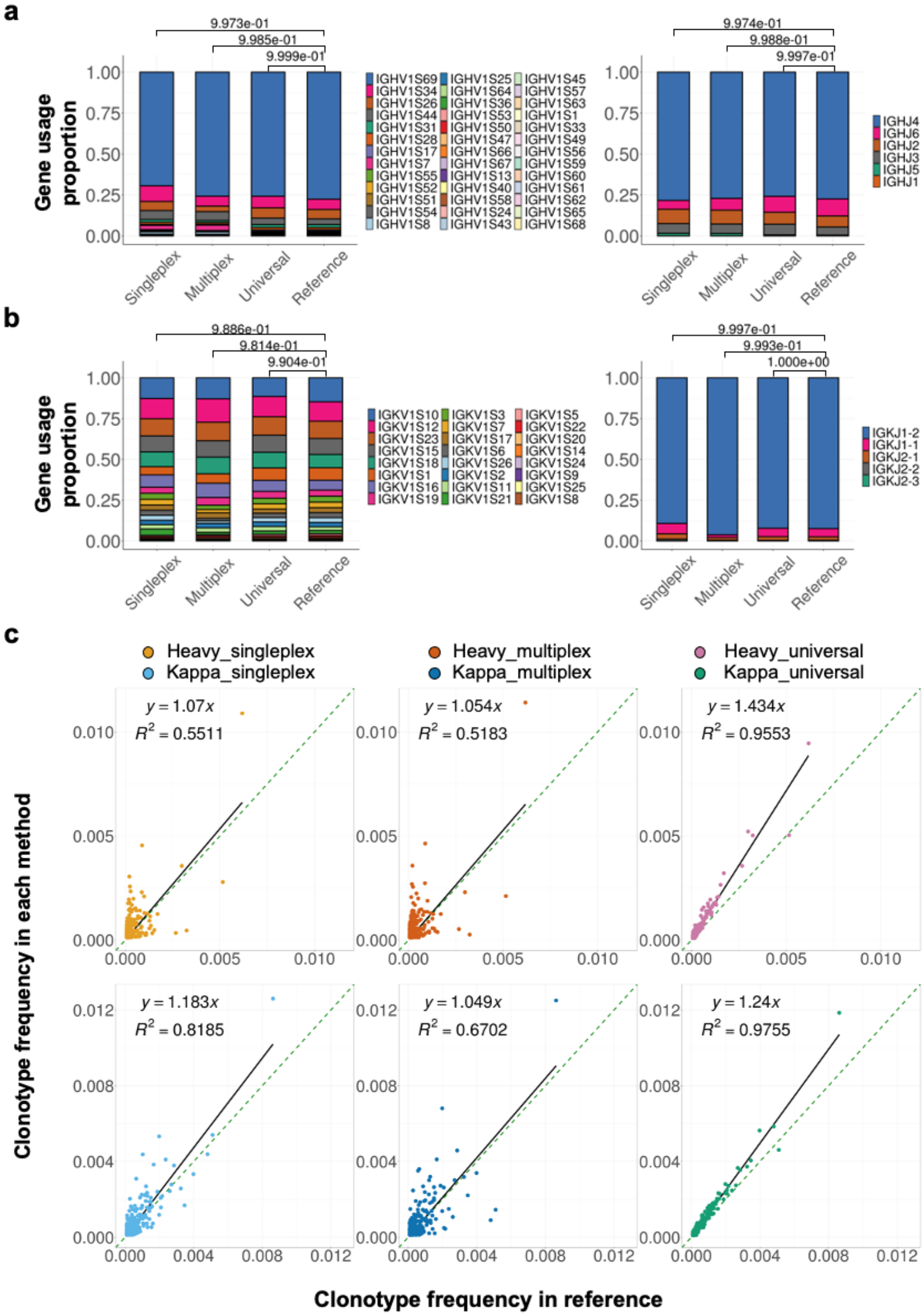
a
9.973e-01
9.985e-01
9.999e-01
Gene usage proportion
1.00
0.75
0.50
0.25
0.00
Singleplex
Multiplex
Universal
Reference
IGHV1S69
IGHV1S34
IGHV1S26
IGHV1S44
IGHV1S31
IGHV1S28
IGHV1S17
IGHV1S7
IGHV1S55
IGHV1S52
IGHV1S51
IGHV1S54
IGHV1S8
IGHV1S25
IGHV1S64
IGHV1S36
IGHV1S53
IGHV1S50
IGHV1S47
IGHV1S66
IGHV1S67
IGHV1S13
IGHV1S40
IGHV1S58
IGHV1S24
IGHV1S43
IGHV1S45
IGHV1S57
IGHV1S63
IGHV1S1
IGHV1S33
IGHV1S49
IGHV1S56
IGHV1S59
IGHV1S60
IGHV1S61
IGHV1S62
IGHV1S65
IGHV1S68
9.974e-01
9.988e-01
9.997e-01
1.00
0.75
0.50
0.25
0.00
Singleplex
Multiplex
Universal
Reference
IGHJ4
IGHJ6
IGHJ2
IGHJ3
IGHJ5
IGHJ1
b
9.886e-01
9.814e-01
9.904e-01
Gene usage proportion
1.00
0.75
0.50
0.25
0.00
Singleplex
Multiplex
Universal
Reference
IGKV1S10
IGKV1S12
IGKV1S23
IGKV1S15
IGKV1S18
IGKV1S1
IGKV1S16
IGKV1S19
IGKV1S3
IGKV1S7
IGKV1S17
IGKV1S6
IGKV1S26
IGKV1S2
IGKV1S11
IGKV1S21
IGKV1S5
IGKV1S22
IGKV1S20
IGKV1S14
IGKV1S24
IGKV1S9
IGKV1S25
IGKV1S8
9.997e-01
9.993e-01
1.000e+00
1.00
0.75
0.50
0.25
0.00
Singleplex
Multiplex
Universal
Reference
IGKJ1-2
IGKJ1-1
IGKJ2-1
IGKJ2-2
IGKJ2-3
c
Heavy_singleplex
Kappa_singleplex
Heavy_multiplex
Kappa_multiplex
Heavy_universal
Kappa_universal
$y = 1.07x$
$R^2 = 0.5511$
$y = 1.054x$
$R^2 = 0.5183$
$y = 1.434x$
$R^2 = 0.9553$
0.010
0.005
0.000
0.000
0.005
0.010
$y = 1.183x$
$R^2 = 0.8185$
$y = 1.049x$
$R^2 = 0.6702$
$y = 1.24x$
$R^2 = 0.9755$
0.012
0.008
0.004
0.000
0.000
0.004
0.008
0.012
Clonotype frequency in each method
Clonotype frequency in reference

Figure 13. Immunoglobin gene usage and clonotype abundance correlation. (a-b) Gene usage proportion of the libraries for the heavy (a) and kappa (b) light chain. V gene (left panel) and J gene (right panel) usage was calculated and compared with the reference data of B cell RNA sample. The similarity was calculated using cosine similarity, which is displayed on top of the figures. Higher cosine similarity values correspond to greater similarity between two libraries. (c) The correlation in frequency of overlapping clonotypes with reference data. To measure the correlation, the regression line and R^2 values were calculated. The green dotted line represents the y=x line. The black line denotes the regression line.

Table 2. Overlapping clonotypes between reference data and the libraries. The number of overlapping clonotypes were counted for top 1,000 and whole clonotypes in both libraries (reference library and the library prepared by each method). The rank of the clonotypes was defined according to the clonal frequency, and if the clonal frequencies of clonotypes were same, rank of the clonotypes were randomly selected.

	Chain type	The number of clonotypes in reference data	Amplification method	The number of clonotypes in each library	The number of overlapping clonotypes	Percentage of overlapping clonotypes (%)
Top 1,000	Heavy chain	1,000	Singleplex	1,000	322	32.2
			Multiplex		329	32.9
			Universal		723	72.3
	Kappa chain		Singleplex		502	50.2
			Multiplex		491	49.1
			Universal		765	76.5
Whole	Heavy chain	56,483	Singleplex	67,200	8,235	14.6
			Multiplex	69,057	8,465	15.0
			Universal	109,047	16,743	29.6
	Kappa chain	40,885	Singleplex	66,428	11,596	28.4
			Multiplex	73,361	11,670	28.5
			Universal	150,884	20,787	50.8

3.4. Discussion

NGS technology has been revolutionizing the path for antibody discovery. It allowed us to analyze BCR repertoires of hosts in depth with great accuracy and provided the information about the isotype frequency, V/J gene usage, accumulated somatic mutation, and foremost, the frequency of individual V_H and V_L sequences [83]. These data are powerful enough to enable the prediction of antigen-binding clonotypes in the repertoire space solely through *in silico* analysis in a controlled situation. In our prior study, we showed that after chronological monitoring on the BCR repertoire of animals undergoing immunization, antigen-reactive V_H clonotypes could be successfully predicted by analyzing the degree of somatic hypermutation and clonotype expansion [91]. NGS analysis on the BCR repertoire of patients affected with viral infection, autoimmune disease, and immunogenic types of cancers revealed the presence of stereotypic Ig clonotypes [92-96], similar to the stereotypic neutralizing V_H clonotypes among patients with COVID-19 that we recently discovered [83]. The presence of these shared Ig clonotypes provided another convenient path for the identification of antibody clones reactive to viral antigens, autoantigens, or tumor-associated antigens.

Single B cell sequencing, combined with antigen-guided B cell selection, helped to determine the frequency of antigen-reactive B cells in peripheral blood [92,97]. Among patients with viral disease, B cells reactive to viral antigen could be relatively rare [97]. To heighten the difficulty of antibody discovery, the B cell clones encoding antibodies with desirable characteristics like virus-neutralizing activity could be even scarcer. For example, the frequency of neutralizing clonotypes was extremely low (0.0004-0.0064%) in the V_H repertoire of patients with COVID-19 [83]. It is well known that immunodominant decoy epitopes are provided by a wide variety of viruses, including human immunodeficiency virus [98], feline immunodeficiency virus [99], hepatitis C [100], foot and mouth disease [101], Middle East respiratory syndrome coronavirus [102], severe fever with thrombocytopenia syndrome virus [103], porcine reproductive and respiratory syndrome virus [104], murine gamma herpesvirus 68 [105], and porcine circovirus type 2 [106], to induce predominantly non-neutralizing antibodies and evade the host immune response.

Antibody display library technology is still one of the most frequently employed high-throughput platforms in antibody discovery from immune hosts. However, to achieve the successful acquisition of rare functional antibodies, the diversity of the BCR repertoire in an RNA sample should be precisely replicated in the antibody display library. Conventionally, for the construction of the antibody display library, gene-specific primers homologous to V and J genes have been employed to amplify V_H and V_L genes. During the PCR amplification of V_H and V_L genes, it was expected that skewing the BCR repertoire occurred, but its extent and characteristics remained un-answered. In prior studies, we established the NGS method to analyze the Ig repertoire of antibody display libraries [64,107]. Using this method, we constructed an *in silico* repertoire of V_H and V_K libraries amplified from B cells of antigen-immunized rabbits through either singleplex or multiplex PCR amplification using gene-specific primers. As expected, rare V_H and V_K clonotypes were preferentially lost during the amplification. In addition, clonotypes with moderate and dominant presence in BCR library tended to be overrepresented in the antibody display library, which limited the fraction of overlapping clonotypes between V_H (~15.0%) and V_K (~28.5%) libraries and the RNA sample of B cells.

To reduce this distortion, we devised a new way to amplify V_H and V_K genes. Double-stranded cDNA was synthesized with primers harboring both a V/J gene-specific sequence and a universal sequence, and PCR amplification was subsequently performed using primers with universal sequences. This universal primer amplification method greatly reduced the diminishment of rare BCR clonotypes (species richness value increased 1.58-2.27 fold) and increased the overlapping clones between V_H (29.6%) and V_K (56.8%) libraries and the RNA sample of B cells. Our study was limited to the rabbit BCR library. Nevertheless, considering that the complexity of Ig genes is similar in other species such as mice, humans, monkeys, and alpacas

(Table 3), our universal primer PCR strategy is likely to be applied with minor modifications.

Table 3. Number of functional immunoglobulin gene in human, mouse rabbit and monkey identified from IMGT database.

Gene type	Human *Homo sapiens*	Mouse *Mus musculus*	Rabbit *Oryctolagus cuniculus*	Monkey *Macaca mulatta*	Alpaca *Vicugna pacos*
IGHV	251	302	39	87	73
IGHJ	12	8	11	9	6
IGKV	64	120	26	83	-
IGKJ	4	8	8	4	-
IGLV	69	5	20	85	-
IGLJ	6	3	2	5	-

IGHV: immunoglobulin heavy chain variable; IGHJ: immunoglobulin heavy chain joining; IGKV: immunoglobulin kappa chain variable; IGKJ: immunoglobulin kappa chain joining; IGLV: immunoglobulin lambda chain variable; IGLJ: immunoglobulin lambda chain joining.

3.5. Methods

Amplification of VH and VK genes

A New Zealand white rabbit (*Oryctolagus cuniculus*) was immunized and then boosted two times with 10 μg recombinant human HGFR/c-MET Fc chimera His-tag protein (358-MT, R&D Systems). One week after the final boosting, total RNA was isolated from the spleen, bone marrow, and peripheral blood mononuclear cells using TRIzol reagent (15596018; Invitrogen). Then, cDNA was synthesized using 1 μg total RNA and a SuperScript Ⅳ first-strand cDNA synthesis kit with oligo dT priming (18091050; Invitrogen). From the cDNA, V_H and V_K amplicons were prepared using singleplex and multiplex PCR. We used rabbit Ig gene-specific primer sets consisting of 14 heavy variable, 5 heavy junctional, 14 kappa variable, and 10 kappa junctional segments, which were rationally designed using the IMGT database by Peng et al. [81] (Table 4). Singleplex amplifications were performed with individual primers in separate reaction tubes. Multiplex amplification was performed using a single PCR reaction with a mixture of multiple primers in an equimolar manner. PCR was performed with KAPA HiFi HotStart DNA polymerase (KK2502; Kappa Bioscience) using forward and reverse primers (95℃ for 3 min, 25 cycles of 98℃ for 30 s, 60℃ for 30 s, 72℃ for 1 min, and 72℃ for 5 min).

For the universal PCR primer library, double-stranded cDNA was synthesized using multiple rabbit Ig gene-specific primers with additional universal sequences, and multicycle amplification was conducted with universal amplification primers (Supplementary Table 3). First-strand cDNA was synthesized using 1 μg total RNA and a SuperScript Ⅳ first-strand cDNA synthesis kit with rabbit J gene-specific reverse primers with an additional restriction or linker sequence, a P7 sequence, and the UMI barcode [80]. First-strand cDNA was purified using AMPure XP beads (A63881; Beckman Coulter) following the instruction provided by the supplier. Then second-strand cDNA was synthesized using KAPA HiFi HotStart DNA polymerase (KK2502; Kappa Bioscience) with rabbit V gene-specific forward primers with an additional restriction or linker sequence and a P5 sequence (98℃ for 4 min, 60℃ for 1 min, 72℃ for 5 min). Double-stranded cDNA was purified using AMPure XP beads (A63881; Beckman Coulter) and subjected to PCR amplification with KAPA HiFi HotStart DNA polymerase (KK2502; Kappa Bioscience) using two universal primers containing Illumina adapters and index sequences (95℃ for 3 min, 25 cycles of 95℃ for 30 s, 65℃ for 30 s, 72℃ for 1 min, and 72℃ for 5 min).

NGS analysis

From the singleplex and multiplex PCR, we prepared 1 μg gel-purified PCR amplicons. Using the purified amplicons, adapter ligation was performed using the NEBNext Multiplex Oligos for Illumina kit (New England Biolabs) following the manufacturer's protocol [91]. Final products were purified using AMPure XP beads (A63881; Beckman Coulter) and submitted to a quality control procedure on TapeStation 2200 (Agilent Technologies) as previously described [83]. Libraries with a single peak of the correct sequence length were subjected to NGS analysis using the MiSeq platform (Illumina Inc.) with 2 × 300 paired-end run mode. From the universal PCR amplification with sample indexing primers, we obtained gel-purified PCR amplicons. Final NGS libraries that passed the quality check were subjected to NGS analysis using the MiSeq platform as described above. To prepare the reference data to accurately reproduce the original BCR repertoire, we constructed and analyzed NGS libraries from distinct RNA input (identical RNA composition used in singleplex, multiplex and universal amplification) with the universal PCR primer amplification method for further UMI-based error correction. We uploaded the sequence data to the National Center for Biotechnology Information (SRA accession number: PRJNA700634).

Preprocessing of raw reads

Raw paired FASTQ files were merged using the PEAR software [108] with default parameters. Merged FASTQ files were quality filtered using an in-house Python 3.6 script with a Q20P95 option, which extracted the reads if more than 95% of bases had a Phred quality score of more than 20. All reads containing more than one N base were excluded. To enhance the validity of the data, a computational error correction process was applied. First, primer sequences were recognized from each read and unrecognized reads were excluded. Primer recognition sites were limited at the edge of reads having a length of 100 bp. Primer sequences were recognized using the BLAST program [109], allowing one mismatch of alignment except for the 6-bp region at the 3′ end of primer sequences. For eliminating the artifacts induced by synthetic errors of primers, primer sequences of each read were trimmed. The end position of the trimming region was determined as the 3′ end position of the primer binding. For the data of singleplex and multiplex gene-specific primer libraries and universal PCR primer libraries, error correction was performed using the MiXCR methodology [110], which corrects errors based on hierarchical clustering. For the reference data, error correction based on the UMI was conducted as previously described [83].

Functional reads filtration and clonotyping

Functional reads were defined as those satisfying following conditions: (1) reads were in-frame without the stop codon and frame shift when translated into amino acid sequences, (2) V (variable) and J (joining) genes were annotated, and (3) complementary determining regions (CDRs; CDR1, CDR2, and CDR3) were extracted without the stop codon and frame shift. The V and J genes and CDRs of each error-corrected read were obtained using the IgBLAST tool [111], with the Ig germline database of the New Zealand white rabbit acquired from the IMGT database [66]. A clonotype was defined as a group of sequences sharing identical V and J genes and encoding identical CDRs at the amino acid level.

Extraction of antibody library features for comparative analysis

Sample coverage and clonotype diversity of six NGS data (singleplex gene-specific PCR primer V_H, multiplex gene-specific PCR primer V_H, universal PCR primer V_H, singleplex gene-specific PCR primer V_K, multiplex gene specific PCR primer V_K, and universal PCR primer V_K) were calculated using the iNEXT R package (version 2017-04-02) [112]. The clonotype diversity was calculated using the iNEXT() function with multiple q values (from 0 to 2), which specify the diversity orders of Hill numbers that denote species richness ($q = 0$), Shannon diversity ($q = 1$), and Simpson diversity ($q = 2$). The V/J gene usage of each library was calculated by summing the clonotype frequency of these genes. The cosine similarity was used for quantifying the similarity of gene usage between the reference data and others, as previously used [113]. The identical clonotypes found both in the reference data and the prepared libraries were extracted, which were denoted as overlapping clonotypes, and the ratio of overlapping clonotypes were quantified. The correlation of clonotype frequency was measured by calculating a regression line of zero y-intercept and the coefficient of determination (R^2, R squared). All statistical analyses except for sample coverage were applied to the clonotypes.

Table 4. List of primers targeting rabbit immunoglobulin V, J genes.

Primers used for singleplex and multiplex PCR

Name	Sequence / Structure	Procedure
Conv_HV1	[GGTGGTTCCTCTAGATCTTCC][CAGTCGGTGGAGGAGTCCAGG]	Multi-cycle PCR
Conv_HV2	[GGTGGTTCCTCTAGATCTTCC][CAGTCGGTGGAGGAGTCCGGG]	Multi-cycle PCR
Conv_HV3	[GGTGGTTCCTCTAGATCTTCC][CAGTCAGTGAAGGAGTCCGAG]	Multi-cycle PCR
Conv_HV4	[GGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGAGGAGTCCGGG]	Multi-cycle PCR
Conv_HV5	[GGTGGTTCCTCTAGATCTTCC][CAGTCGTTGGAGGAGTCCGGG]	Multi-cycle PCR
Conv_HV6	[GGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCTGGAGGAGTCCGGG]	Multi-cycle PCR
Conv_HV7	[GGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCTGAAGGAGTCCGG]	Multi-cycle PCR
Conv_HV8	[GGTGGTTCCTCTAGATCTTCC][CAGAAGCAGCTGGTGGAGTCCGG]	Multi-cycle PCR
Conv_HV9	[GGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCTGGTGGAGTCCGG]	Multi-cycle PCR
Conv_HV10	[GGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCAGAAGGAGTCCGGG]	Multi-cycle PCR
Conv_HV11	[GGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGAGGAGTCCAGG]	Multi-cycle PCR
Conv_HV12	[GGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGGGGAGTCCAGG]	Multi-cycle PCR
Conv_HV13	[GGTGGTTCCTCTAGATCTTCC][CAGACAGTGAAGGAGTCCGAG]	Multi-cycle PCR
Conv_HV14	[GGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGAGGAATTCGGG]	Multi-cycle PCR
Conv_HV_structure	[Linker partial][V gene specific region]	
Conv_HJ1	[CCTGGCCGGCCTGGCCACTAGT][TGAAGAGACGGTGACCAGGGTGCC]	Multi-cycle PCR
Conv_HJ2	[CCTGGCCGGCCTGGCCACTAGT][TGAAGAGATGGTGACCAGGGTGCC]	Multi-cycle PCR
Conv_HJ3	[CCTGGCCGGCCTGGCCACTAGT][TGAGGAGACGGTGACCAGGGTGCC]	Multi-cycle PCR
Conv_HJ4	[CCTGGCCGGCCTGGCCACTAGT][TGAGGAGATGGTGACCAGGGTGCC]	Multi-cycle PCR
Conv_HJ5	[CCTGGCCGGCCTGGCCACTAGT][TGAAGAGACGGTGACGAGGGTCCC]	Multi-cycle PCR
Conv_HJ_structure	[Sfi I restriction site][J gene specific region]	
Conv_KV1	[GGGCCCAGGCGGCC][GCCGCCGTGCTGACCCAGACT]	Multi-cycle PCR
Conv_KV2	[GGGCCCAGGCGGCC][GCCCAAGTGCTGACCCAGACT]	Multi-cycle PCR
Conv_KV3	[GGGCCCAGGCGGCC][GCCCTTGTGATGACCCAGACT]	Multi-cycle PCR
Conv_KV4	[GGGCCCAGGCGGCC][GACCCTATGCTGACCCAGACT]	Multi-cycle PCR
Conv_KV5	[GGGCCCAGGCGGCC][GATGTCGTGATGACCCAGACT]	Multi-cycle PCR
Conv_KV6	[GGGCCCAGGCGGCC][GACCCTGTGCTGACCCAGAC	Multi-cycle PCR

	T]	
Conv_KV7	[GGGCCCAGGCGGCC][TATGTCATGATGACCCAGACT]	Multi-cycle PCR
Conv_KV8	[GGGCCCAGGCGGCC][GCCGCCGTGATGACCCAGACT]	Multi-cycle PCR
Conv_KV9	[GGGCCCAGGCGGCC][GCCCAAGGGCCAACCCAGACT]	Multi-cycle PCR
Conv_KV10	[GGGCCCAGGCGGCC][GCCGTCGTGCTGACCCAGACT]	Multi-cycle PCR
Conv_KV11	[GGGCCCAGGCGGCC][GCCATCAAAATGACCCAGACT]	Multi-cycle PCR
Conv_KV12	[GGGCCCAGGCGGCC][GACCCTGTGATGACCCAGACT]	Multi-cycle PCR
Conv_KV13	[GGGCCCAGGCGGCC][GATGGCGTGATGACCCAGACT]	Multi-cycle PCR
Conv_KV14	[GGGCCCAGGCGGCC][GACATTGTGCTGACCCAGACT]	Multi-cycle PCR
Conv_KV_structure	[Sfi I restriction site][V gene specific region]	
Conv_KJ1	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGATTTCCACATTGGTGCC]	Multi-cycle PCR
Conv_KJ2	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGATTTCCACCTTGGTGCC]	Multi-cycle PCR
Conv_KJ3	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGATCTCCACCTTGGTCCC]	Multi-cycle PCR
Conv_KJ4	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGATCTCCACCTTGGTTCC]	Multi-cycle PCR
Conv_KJ5	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGATCTCCAGCTTGGTCCC]	Multi-cycle PCR
Conv_KJ6	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGATCTCCAGCTTGGTTCC]	Multi-cycle PCR
Conv_KJ7	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGACCACCACCTCGGTCCC]	Multi-cycle PCR
Conv_KJ8	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTTGACGACCACCTCGGTCCC]	Multi-cycle PCR
Conv_KJ9	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TAGGATCTCCAGCTCGGTCCC]	Multi-cycle PCR
Conv_KJ10	[GGAAGATCTAGAGGAACCACCCCCACCACCGCCCGAGCCACCGCCACCAGAGGA][TTCGACGACCACCTTGGTCCC]	Multi-cycle PCR
Conv_KJ_structure	[Linker partial][J gene specific region]	

(continued)

Name	Sequence	Procedure
Alt_HV1	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGGTGGAGGAGTCCAG]	2nd strand cDNA synthesis
Alt_HV2	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGGTGGAGGAGTCCGGG]	2nd strand cDNA synthesis
Alt_HV3	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCAGTGAAGGAGTCCGAG]	2nd strand cDNA synthesis
Alt_HV4	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGAGGAGTCCGGG]	2nd strand cDNA synthesis
Alt_HV5	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGTTGGAGGAGTCCGGG]	2nd strand cDNA synthesis
Alt_HV6	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCTGGAGGAGTCCGGG]	2nd strand cDNA synthesis
Alt_HV7	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCTGAAGGAGTCCGG]	2nd strand cDNA synthesis
Alt_HV8	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGAAGCAGCTGGTGGAGTCCGG]	2nd strand cDNA synthesis
Alt_HV9	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCTGGTGGAGTCCGG]	2nd strand cDNA synthesis
Alt_HV10	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGGAGCAGCAGAAGGAGTCCGGG]	2nd strand cDNA synthesis
Alt_HV11	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGAGGAGTCCAGG]	2nd strand cDNA synthesis
Alt_HV12	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGGGGAGTCCAGG]	2nd strand cDNA synthesis
Alt_HV13	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGACAGTGAAGGAGTCCGAG]	2nd strand cDNA synthesis
Alt_HV14	[CACGACGCTCTTCCGATCT][GCGGTGGTGGGGGTGGTTCCTCTAGATCTTCC][CAGTCGCTGGAGGAATTCGGG]	2nd strand cDNA synthesis
Alt_HV_structure	[P5 illumina adapter partial][linker partial][V gene specific region]	
Alt_HJ1	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCTGGCCGGCCTGGCCACTAGT][TGAAGAGACGGTGACCAGGGTGCC]	Reverse transcription
Alt_HJ2	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCTGGCCGGCCTGGCCACTAGT][TGAAGAGATGGTGACCAGGGTGCC]	Reverse transcription
Alt_HJ3	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCTGGCCGGCCTGGCCACTAGT][TGAGGAGACGGTGACCAGGGTGCC]	Reverse transcription
Alt_HJ4	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCTGGCCGGCCTGGCCACTAGT][TGAGGAGATGGTGACCAGGGTGCC]	Reverse transcription

Alt_HJ5	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCTGGCCGGCCTGGCCACTAGT][TGAAGAGACGGTGACGAGGGTCCC]	Reverse transcription
Alt_HJ_structure	[P7 illumina adapter partial][UMI barcode][Sfi I restriction site][J gene specific region]	
Alt_KV1	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCGCCGTGCTGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV2	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCCAAGTGCTGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV3	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCCTTGTGATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV4	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GACCCTATGCTGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV5	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GATGTCGTGATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV6	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GACCCTGTGCTGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV7	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][TATGTCATGATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV8	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCGCCGTGATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV9	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCCAAGGGCCAACCCAGACT]	2nd strand cDNA synthesis
Alt_KV10	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCGTCGTGCTGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV11	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GCCATCAAAATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV12	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GACCCTGTGATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV13	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GATGGCGTGATGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV14	[CACGACGCTCTTCCGATCT][GGGCCCAGGCGGCC][GACATTGTGCTGACCCAGACT]	2nd strand cDNA synthesis
Alt_KV_structure	[P5 illumina adapter partial][Sfi I restriction site][V gene specific region]	
Alt_KJ1	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGATTTCCACATTGGTGCC]	Reverse transcription
Alt_KJ2	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGATTTCCACCTTGGTGCC]	Reverse transcription
Alt_KJ3	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGATCTCCACCTTGGTCCC]	Reverse transcription
Alt_KJ4	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGATCTCCACCTTGGTTCC]	Reverse transcription
Alt_KJ5	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGATCTCCAGCTTGGTCCC]	Reverse transcription
Alt_KJ6	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGATCTCCAGCTTGGTTCC]	Reverse transcription
Alt_KJ7	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGACCACCACCTCGGTCCC]	Reverse transcription
Alt_KJ8	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTTGACGACCACCTCGGTCCC]	Reverse transcription

Alt_KJ9	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TAGGATCTCCAGCTCGGTCCC]	Reverse transcription
Alt_KJ10	[ACGTGTGCTCTTCCGATCT][NNNNTNNNNTNNNN][CCGAGCCACCGCCACCAGAGGA][TTCGACGACCACCTTGGTCCC]	Reverse transcription
Alt_KJ_structure	[P7 illumina adapter partial][UMI barcode][linker partial][J gene specific region]	

(continued)

Name	Sequence	Procedure
Uni_forward_p5	CACGACGCTCTTCCGATCT	Multi-cycle PCR
Uni_reverse_p7	ACGTGTGCTCTTCCGATCT	Multi-cycle PCR
Linker_adapter	CTCTGGTGGCGGTGGCTCGGGCGGTGGTGGGGGTGGTTC	Overlap extension PCR
Overlap_forward	GAGGCGGGGCCCAGGCGGCCGAGC	Overlap extension PCR
Overlap_reverse	GAGCCTGGCCGGCCTGGCCACTAGTG	Overlap extension PCR
Uni_forward_p5	CACGACGCTCTTCCGATCT	Multi-cycle PCR
Uni_reverse_p7	ACGTGTGCTCTTCCGATCT	Multi-cycle PCR
Linker_adapter	CTCTGGTGGCGGTGGCTCGGGCGGTGGTGGGGGTGGTTC	Overlap extension PCR

4. *In silico* identification of target specific antibodies by high-throughput antibody repertoire analysis and machine learning

This is an author produced adaptation of a peer reviewed article published in

Experimental & Molecular Medicine (2017). doi:0.1038/emm.2017.22.
[Yang W, Yoon A, Lee S, Kim S, Han J & Chung J]

Biomolecules (2020). doi:10.3390/biom10030421.
[Yoo DK*, Lee SR*, Jung Y, Han H, Lee HK, Han J, Kim S, Chae J, Ryu T & Chung J]

* These authors contributed equally.

4.1. Abstract

c-Met is a promising target in cancer therapy for its intrinsic oncogenic properties. However, there are currently no c-Met-specific inhibitors available in the clinic. Antibodies blocking the interaction with its only known ligand, hepatocyte growth factor, and/or inducing receptor internalization have been clinically tested. To explore other therapeutic antibody mechanisms like Fc-mediated effector function, bispecific T cell engagement, and chimeric antigen T cell receptors, a diverse panel of antibodies is essential. We prepared a chicken immune scFv library, performed four rounds of bio-panning, obtained 641 clones using a high-throughput clonal retrieval system (TrueRepertoire^{TM}, TR), and found 149 antigen-reactive scFv clones. We also prepared phagemid DNA before the start of bio-panning (round 0) and, after each round of bio-panning (round 1-4), performed next-generation sequencing of these five

sets of phagemid DNA, and identified 860,207 HCDR3 clonotypes and 443,292 LCDR3 clonotypes along with their clonal abundance data. We then established a TR data set consisting of antigen reactivity for scFv clones found in TR analysis and the clonal abundance of their HCDR3 and LCDR3 clonotypes in five sets of phagemid DNA. Using the TR data set, a random forest machine learning algorithm was trained to predict the binding properties of in silico HCDR3 and LCDR3 clonotypes. Subsequently, we synthesized 40 HCDR3 and 40 LCDR3 clonotypes predicted to be antigen reactive (AR) and constructed a phage-displayed scFv library called the AR library. In parallel, we also prepared an antigen non-reactive (NR) library using 10 HCDR3 and 10 LCDR3 clonotypes predicted to be NR. After a single round of bio-panning, we screened 96 randomly-selected phage clones from the AR library and found out 14 AR scFv clones consisting of 5 HCDR3 and 11 LCDR3 AR clonotypes. We also screened 96 randomly-selected phage clones from the NR library, but did not identify any AR clones. In summary, machine learning algorithms can provide a method for identifying AR antibodies, which allows for the characterization of diverse antibody libraries inaccessible by traditional methods.

4.2. Introduction

The mesenchymal-epithelial transition factor (c-Met) and its ligand hepatocyte growth factor (HGF) are well-known oncogenic drivers of tumorigenesis[114]. Numerous clinical observations have demonstrated that c-Met overexpression or gene alterations play a key role in both oncogenesis and the development of drug resistance across multiple cancer types[115-118]. Furthermore, recent research suggests that the HGF-c-Met axis limits the efficacy of cancer immunotherapy by modulating immune cell function and the expression of programmed cell death ligand 1 (PD-L1)[119-122]. Despite efforts to inhibit the HGF-c-Met axis including antibodies against c-Met or HGF, c-Met tyrosine kinase inhibitors, and more, no therapeutic agent specific to the HGF-c-Met axis is clinically available. Currently, two anti-HGF antibodies, including YYB-101 previously discovered by our group, are under clinical trials (NCT02499224)[123]. However, no antibodies are under development against c-Met after the failure of onartuzumab in clinical trials[124].

Based on rapid advances in next-generation sequencing (NGS) technology, various methodologies for analyzing NGS data have been developed to decode the antibody repertoire from diverse sources such as the natural B cell receptor of animals and humans as well as recombinant antibody libraries that can be synthetically designed and constructed[125-127]. Furthermore, combining surface display technology and NGS analysis offers synergistic advantages in identifying antigen-reactive clones in silico over the laborious in vitro screening process, which is frequently overwhelmed by dominant antibody clones[48]. Traditional bio-panning methodologies are biased towards the excessive enrichment of dominant clones with significant suppression of antibody diversity. Consequently, this approach could lead to the omission of potential antigen-reactive (AR) clones with low clonal abundance or their diminishment by unintended selective pressure.

Previously, our group analyzed the enrichment patterns of bio-panned clones by employing NGS technology to predict the antigen binding properties of antibody clones inside different clusters[64]. First, we tracked the clonal abundance of heavy chain complementarity region 3 (HCDR3) through multiple rounds of bio-panning with NGS analysis, and then applied clustering analysis to group HCDR3 clonotypes based on the enrichment pattern. As a result, different clusters (enriched, impoverished, and fluctuated) were generated with the enriched pattern cluster containing a higher frequency of AR scFv (single-chain variable fragment) clones than other clusters. However, due to limitations in retrieving the physical DNA of encoded scFv from a large, diverse number of clones, we were unable to sufficiently observe the binding properties of in silico scFv clones. Recently, we developed a laser and microchip-based high-throughput clonal retrieval system (TrueRepertoire^{TM}, TR) for scFv DNA from the library[63], which allows a much higher number of scFv clones to be obtained and tested for antigen reactivity.

In this study, we established a phage-displayed chicken scFv library after immunization with recombinant c-Met. Four rounds of bio-panning were performed on antigen-conjugated magnetic beads. Through bio-panning, five sets of phagemid DNA (rounds 0–4) were obtained and subjected to NGS analysis using both HiSeq and MiSeq platforms. After the final round of bio-panning, scFv-displayed phage clones were obtained in a high-throughput manner using TR technology, and individual clone reactivity was evaluated by phage enzyme-linked immunosorbent assay (ELISA). From NGS data obtained using the HiSeq platform, HCDR3, and light chain complementarity region 3 (LCDR3) clonotypes were extracted and evaluated for their clonal abundance in phagemid DNA sets from round 0 (before biopanning) to round 4. We then established a data set (TR data set) containing the antigen reactivity of scFv clones retrieved through TR technology and the clonal abundance of their HCDR3 and LCDR3 clonotypes in five sets of phagemid DNA. Using this TR data set, we trained our random forest (RF) machine learning algorithm to predict the binding properties of in silico HCDR3 and LCDR3 clonotypes[128,129].

To test the accuracy of our RF model (Figure 14), we extracted V_H and V_L sequences from MiSeq NGS data, which encompass both RF model-determined AR and antigen non-reactive (NR) HCDR3 or LCDR3 clonotypes and chemically synthesized them. Using these V_H and V_L genes, we established two phage-displayed scFv libraries. The AR library was prepared using V_H and V_L genes encompassing AR HCDR3 and LCDR3 clonotypes, and the NR library was constructed using V_H and V_L genes encompassing NR HCDR3 and LCDR3 clonotypes. After one round of bio-panning on antigen-conjugated magnetic beads, antigen reactivity of phage clones was tested by phage ELISA. From the AR library, we obtained many scFv clones containing AR HCDR3 and LCDR3 clonotypes, while no AR clones were enriched from the NR library.

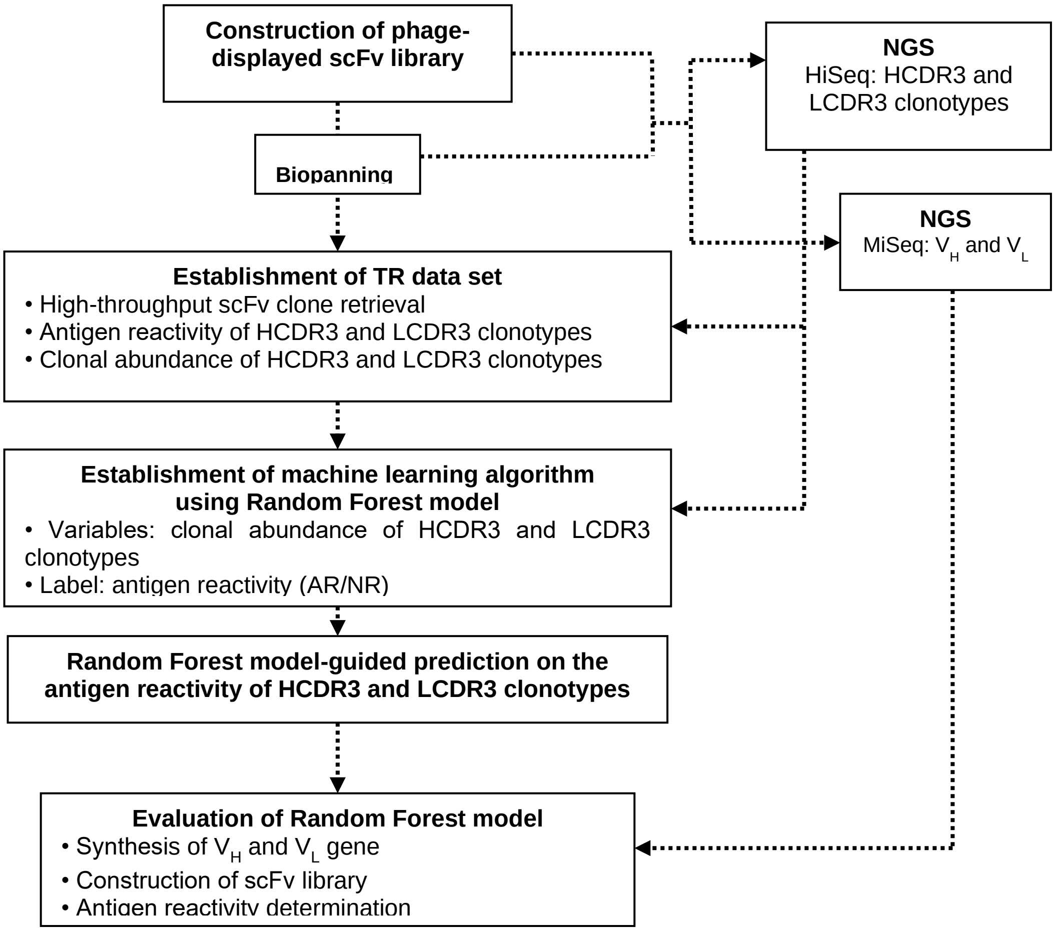
Construction of phage-displayed scFv library
Biopanning
NGS
HiSeq: HCDR3 and LCDR3 clonotypes
NGS
MiSeq: V_H and V_L
Establishment of TR data set
• High-throughput scFv clone retrieval
• Antigen reactivity of HCDR3 and LCDR3 clonotypes
• Clonal abundance of HCDR3 and LCDR3 clonotypes
Establishment of machine learning algorithm using Random Forest model
• Variables: clonal abundance of HCDR3 and LCDR3 clonotypes
• Label: antigen reactivity (AR/NR)
Random Forest model-guided prediction on the antigen reactivity of HCDR3 and LCDR3 clonotypes
Evaluation of Random Forest model
• Synthesis of V_H and V_L gene
• Construction of scFv library
• Antigen reactivity determination

Figure 14. Workflow of the machine learning-guided selection of antigen-reactive HCDR3 and LCDR3 clonotypes with confirmation of their reactivity.

4.3. Results

Generation of antibody library and screening for positive clones using the conventional colony screening method

Using mRNA prepared from spleen, bone marrow, and bursa of Fabricius from three PSA-immunized chickens, we generated scFv libraries with complexities of 6.09×10^{10}, 3.64×10^{10} and 5.16×10^{10} clones, respectively, referred to as chicken libraries 1, 2 and 3. Next, we performed four rounds of bio-panning, rescued phage clones from the output titer plate of the fourth round, and performed a phage enzyme immunoassay to screen for positive clones. A total of 300 clones (100 clones in each library) exhibiting an optical density of >0.3 at 405 nm were considered to be positive, and their scFv gene sequence was determined by Sanger sequencing analysis. We finally obtained 22 clones with unique HCDR3 sequences (Table 5).

Table 5. HCDR3 amino-acid sequences selected using the conventional colony screening method, and binding reactivity measurement of the antibody clones.

Library	Cluster	Sequence of HCDR3	Proportion of NGS (%)	Proportion of conventional method (%)	Binding reactivity (O.D.405 nm)
1	1	DFGSGVGEIDA	3.81	1.04	1.01
		GIESDSDGYMTAEEIDA	0.13	1.04	0.977
	2	AAHSTYIWGGYEAGSIDA	6.49	4.17	0.669
		SAVSSCSSGSCSASWIDA	1.16	2.08	0.873
		TADDGFSCGGYGLCADRIDA	0.39	1.04	0.723
		ESGNGGWITAARIDA	0.08	1.04	0.767
		SSHSTYIWGAYEAGSIDA	0.03	2.08	0.651
	4	APGTGSGYCGIWTYTTAGCIDA	0.03	1.04	0.964
		GRISYICADYDAGCIDA	0.02	5.21	1.063
		SSHSTYIWGGYEAGSIDA	0.01	2.08	0.916
2	2	SSYSDGATVIYNIDA	0.69	1.04	0.87
	3	GRISYICADYDAGCIDA	0.04	6.25	1.063
		AAGSWCAWGTGSCAGSIDA	0.02	5.21	1.067
		AAGSWCAWGTGSCAGNIDA	0.01	1.04	0.985
		TTGGDFYSGIDTAGYIDA	0.01	5.21	0.938
		APGTGSGYCGIWTYTTAGCIDA	0.01	3.13	0.964
3	2	AAGSGYIYSGSAGWIDA	1.07	3.13	0.941
	3	AAGSWCAWGTGSCAGSIDA	0.03	4.17	0.918
		GRISYICADYDAGCIDA	0.02	8.33	1.063
		TTGGDFYSGIDTAGYIDA	0.02	2.08	0.889
		AAGSWCAWGAGSCAGSIDA	0.01	1.04	0.914
		AAGSGYVYSGSAGWIDA	0.01	2.08	1.021

HCDR3; heavy chain complementarity-determining region 3; NGS, next-generation sequencing; O.D., optical density.

Diversity analysis of antibody clones using next-generation sequencing

A total of 15 sets of phagemid DNA (three chicken libraries from bio-panning rounds 0, 1, 2, 3, and 4) were used for NGS analysis. After the NGS experiment, we obtained 60,000–180,000 VH sequences. Raw paired-end nucleotide sequences were merged, filtered, aligned and trimmed by uniformly applying pre-specified criteria to remove low-quality and meaningless short sequences. The numbers of nucleotide sequences remaining after each preprocess are summarized in Table 6; and were used in subsequent analyses. From the NGS results, the total population of VH fragment nucleotides decreased as the bio-panning rounds proceeded. To analyze HCDR3 diversity and frequency, we used HCDR3 sequences existing only in the fourth bio-panning round. clValid predicted that 2–6 clusters would be the most dependable in the HCDR3 sequence count profile data (Table 7). The sequence reads in chicken library 1 showed the maximum Dunn index (0.1048) with 4–6 clusters, and chicken libraries 2 and 3 had maximum Dunn indices with 2–3 clusters. We clustered HCDR3 sequences into 2–6 clusters using hierarchical clustering, and generated heat maps for each cluster to examine the patterns of HCDR3 sequence enrichment and population shift throughout the bio-panning rounds. The pattern of HCDR3 sequence enrichment and population shift in chicken library 1 showed four clear clusters, and the patterns in chicken libraries 2 and 3 showed three clear clusters (Figure 15).

Table 6. Sequence read counts by preprocessing raw sequencing data.

Library	Panning round	Raw sequencing read count	Read count after merging	Read ount of qualified sequences	Read count of disqualified sequences	Read count aligned with HCD R3	Unique nucleotide sequence count
1	R0	664,955	393,749	393,624	125	310,589 (78.9)	205,255
	R1	663,061	377,630	377,484	146	298 474 (79)	198,150
	R2	391,118	229,873	229,773	100	181,430 (78.9)	128,513
	R3	673,875	388,341	388,179	162	314,517 (81)	148,787
	R4	621,174	379,630	379,611	19	334,387 (88.1)	27,141
2	R0	432,274	256,268	256,199	69	193,262 (75.4)	148,862
	R1	661,248	417,426	417,323	103	316,150 (75.7)	221,423
	R2	608,850	363,553	363,460	93	274,100 (75.4)	197,190
	R3	547,353	342,189	342,123	66	289,287 (84.5)	66,545
	R4	455,119	290,741	290,722	19	274,635 (94.5)	22,763
3	R0	616,410	360,830	360,783	47	279,996	164,869

					(77.6)	
R1	608,045	370,090	370,033	57	288,172 (77.9)	167,249
R2	619,731	373,093	373,038	55	290,056 (77.7)	168,084
R3	690,602	419,796	419,757	39	343,996 (81.9)	74,611
R4	568,948	354,314	354,301	13	287,126 (81)	21,884

HCDR3, heavy chain complementarity-determining region 3.

Table 7. Dunn index on hierarchical clustering to estimate optimal number of clusters in scFv nucleotide sequence profile data

	Number of clusters				
	2	3	4	5	6
Library 1	0.0863	0.0723	0.1048	0.1048	0.1048
Library 2	0.2331	0.2331	0.0564	0.0564	0.0845
Library 3	0.1508	0.186	0.1544	0.0893	0.0893

scFv, single-chain variable fragment. Bold numbers indicate the largest Dunn index in each library.

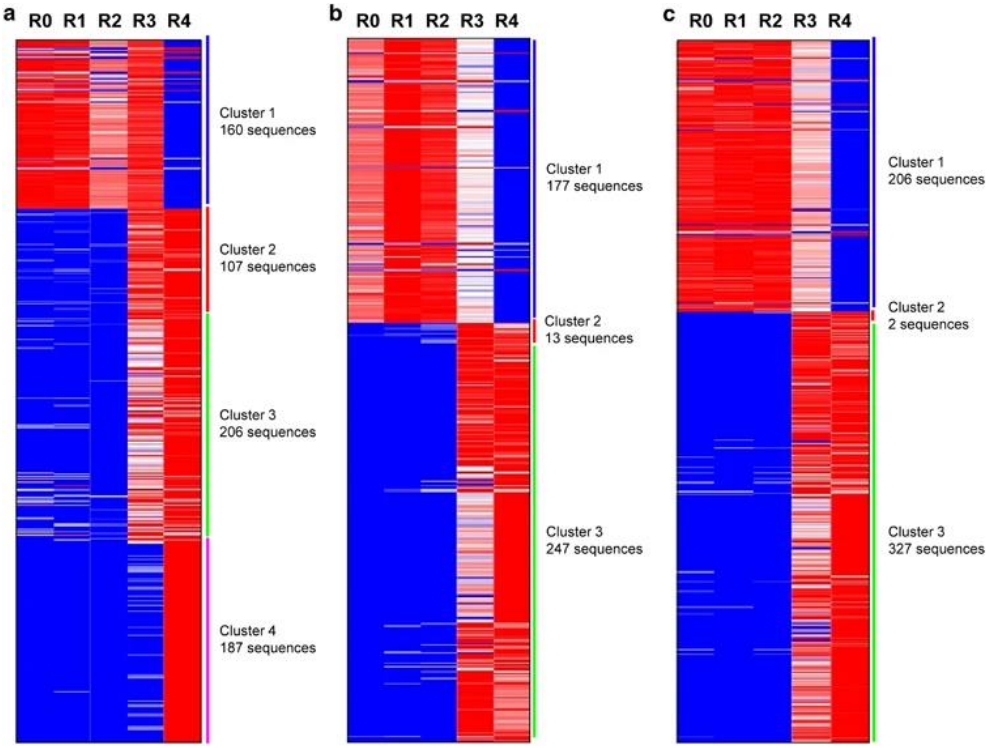

Figure 15. Heat map representing the population of heavy chain complementarity-determining region 3 (HCDR3) sequences in each cluster through bio-panning rounds. Red and blue denote high and low proportions of the HCDR3 sequence, respectively. (a) scFv library 1, (b) scFv library 2 and (c) scFv library 3.

Population shift in HCDR3 sequences throughout bio-panning rounds

The diversity of the antibody clones is represented by the number of HCDR3 sequences that belong to each cluster (Figure 16). The abundance of the HCDR3 sequences in each cluster is represented by heat map color; high and low populations are indicated in red and blue, respectively. HCDR3 sequences in cluster 1 were highly abundant before bio-panning and up to the second bio-panning round. However, there was a sudden impoverishment in rounds 3 and 4 of bio-panning. In contrast, HCDR3 sequences that belonged to clusters 2 and 3 (including cluster 4 of library 1) showed the opposite pattern. Their populations were very low before bio-panning, remained low after the second round of bio-panning, and started to enrich from the third round of bio-panning. The increase continued in the fourth round of bio-panning. This population shift of HCDR3 sequences throughout bio-panning is represented in Figure 16. All 22 HCDR3 sequences in clones found via the conventional colony screening method existed among the HCDR3 sequences obtained from NGS analysis of phagemid DNA prepared after the fourth round of bio-panning (Table 5). Two out of the 22 unique HCDR3 sequences belonged to cluster 1, and the other 20 HCDR3 sequences belonged to clusters 2, 3 or 4.

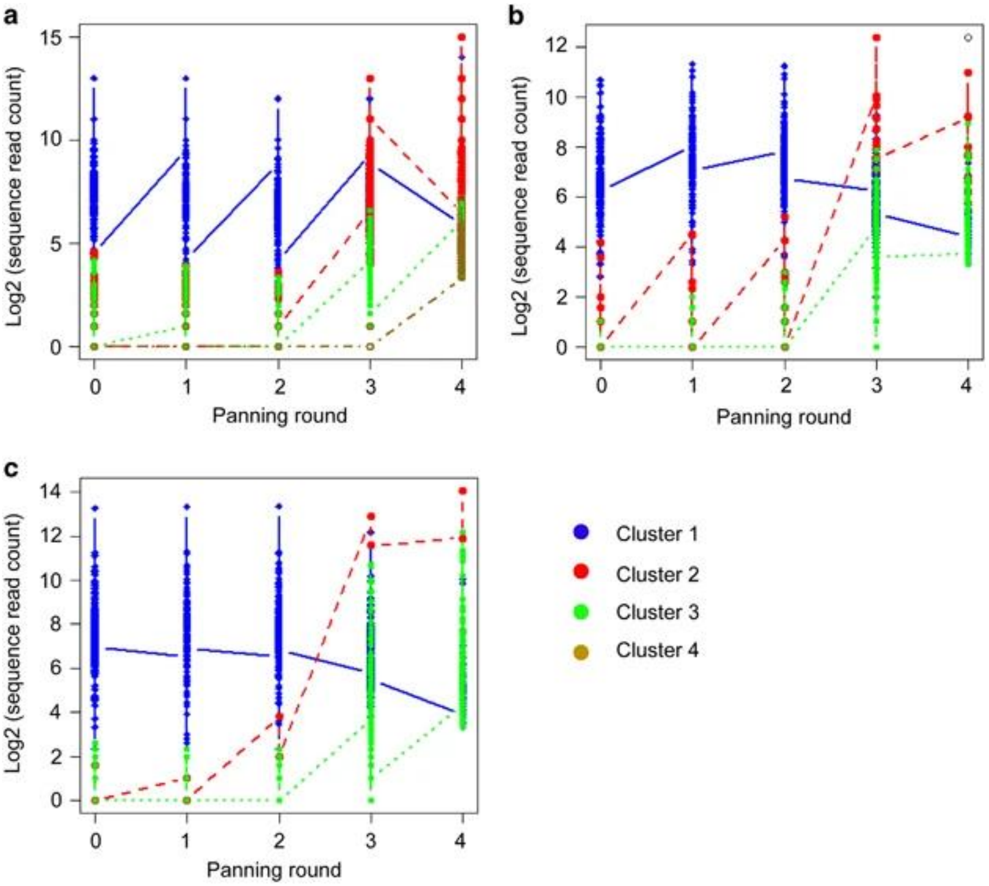

Figure 16. Line graph representing population shifts in HCDR3 sequences through bio-panning rounds. (a) scFv library 1, (b) scFv library 2 and (c) scFv library 3.

Reactivity of scFv clones identified in NGS analysis

For each cluster, 1–5 HCDR3 sequences newly identified from the fourth round of bio-panning via NGS analysis were selected arbitrarily (Table 8). These selected sequences were used to synthesize the primers to retrieve the whole scFv gene from the phagemid DNA. The scFv gene was prepared in two-step linker PCR using the primers and cloned into a phagemid vector (Figure 17). After transformation of the phagemid vector-encoding scFv gene and rescue with helper phage, scFv-displaying phage was used to test their binding reactivity against PSA (Figure 18). In cluster 1, across the three libraries, 12 out of 14 antibody clones (85.7%) had negligible binding reactivity against PSA (O.D.450nm<0.2; Table 4, blue). In contrast, 21 out of 26 antibody clones (80.8%) in clusters 2~4 across the three libraries had significant binding reactivity (O.D.450nm>0.3; Table 4, red). These results imply that antibody clones with low reactivity tend to be impoverished throughout bio-panning (cluster 1), in contrast to the antibody clones with high reactivity, which showed enrichment throughout bio-panning (clusters 2~4).

Table 8. CDR3 amino-acid sequences selected in each cluster from NGS and binding reactivity measurement of antibody clones.

Library	Cluster	HCDR3 Sequence	Proportion of the sequence in R4	O.D.405 nm
1	Cluster 1	GVYSGSPDGYDIDA	0.32%	0.454
		TTCVGSSYCGGENIDA	0.16%	0.173
		GAYSDWGAGFIDA	0.08%	0.161
		DGDSGWGVYLNSAGNIDA	0.03%	0.153
	Cluster 2	YAGSGWTYYSSDVGSIDA	2.16%	0.62
		GVYSASGCCDSIDT	1.93%	1.032
		SAHSTYIWGGYEAGSIDA	1.41%	1.075
		GGGAGYGAPSIDT	1.05%	0.871
		DVYSGLITANTIDA	0.67%	0.639
	Cluster 3	SSHSTYIWGAYEAGCIDA	0.02%	0.757
		RAYGGGYCGCIEDIDA	0.01%	0.323
		AASTWSFYGSAEDIDA	0.01%	0.725
	Cluster 4	APGTGSGYCGIWTYTTAGSIDA	0.04%	0.323
		GRISYICADYEAGSIDA	0.02%	0.407
2	Cluster 1	GAYGHCDGWCAVDSIDT	0.07%	0.175
		AAGSGYCGWGDCIAGSIDA	0.07%	0.167
		GIYGYSGGDYAAAEIDA	0.06%	0.179
		GAGGSCDGGSWCSPGIIDA	0.04%	0.187
		TRGGAGSGWYWYSGIAGIIDA	0.03%	0.18
	Cluster 2	TAGCGPWSYITAGCIDA	0.21%	1.119
		DAAYGYCGTWAGCAGRIDA	0.21%	1.187
		CAYSGCTGGWSTSSIDA	0.20%	1.007
		DVYGCNSYGCPYIGNTIDA	0.09%	1.254
		RAFSGCCDADSIDA	0.07%	0.845
	Cluster 3	SSSGTTYYSSGVISAGGIDA	0.17%	0.167
		GRISYICVDYDAGCIDA	0.07%	0.706
		NAYTSAYITDIDS	0.06%	0.944

		SAYSDSCCAEDIDA	0.04%	0.876
		SAFGGGACCYTAGTIDA	0.03%	0.165
3	Cluster 1	DGSGCGWSAAGCIDA	0.35%	0.16
		AATYSWLHSGIDA	0.29%	0.728
		DGSDCGWSAAGCIDA	0.06%	0.146
		GTGSWCYSGADSIDT	0.06%	0.167
		SAAGYWYAGSIDA	0.05%	0.138
	Cluster 2	TAGGDFYSGVDTAGYIDA	4.79%	1.064
	Cluster 3	GSGYSCWSYAGCIDA	0.66%	1.083
		GRIYYICADYDAGCIDA	0.53%	1.052
		TADSGFGCGGYGLCAAFIDA	0.09%	0.907
		TADIGYCFGGGIGCIDA	0.08%	0.984
		SAGGSYGYRYMDTAAAIDA	0.07%	0.861

HCDR3; heavy chain complementarity-determining region 3; NGS, next-generation sequencing.

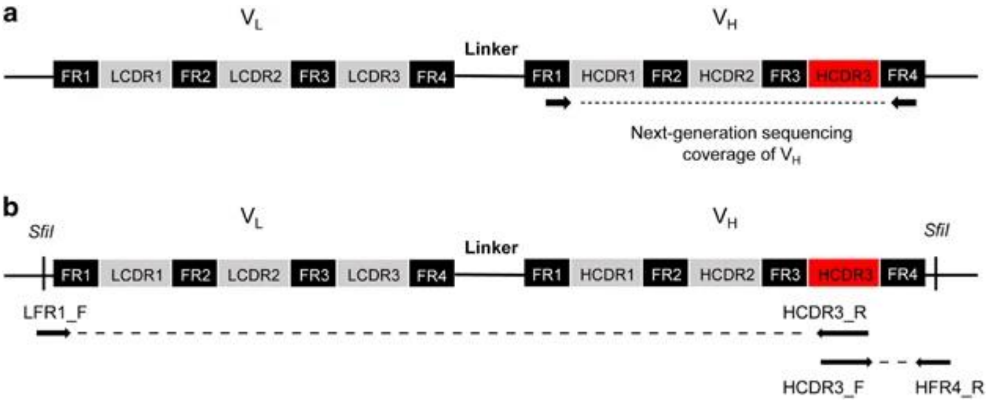

Figure 17. Schematic representation of next-generation sequencing and two-step linker PCR. The structure of scFv gene, CDRs and frameworks of variable regions are indicated by colored boxes. (a) For NGS analysis, most of VH region including HCDR3 was amplified and sequenced using specific primers as described in materials and methods. The sequencing coverage is indicated with dashed lines. (b) To retrieve scFv gene, two-step linker PCR was performed using primers annealing to HCDR3, LFR1 and HFR4. The first step of PCR was performed using LFR1_F and HCDR3_R primers and HCDR3_F and HFR4_R primers. The linker PCR was performed using LFR1_F and HFR4_R primers.

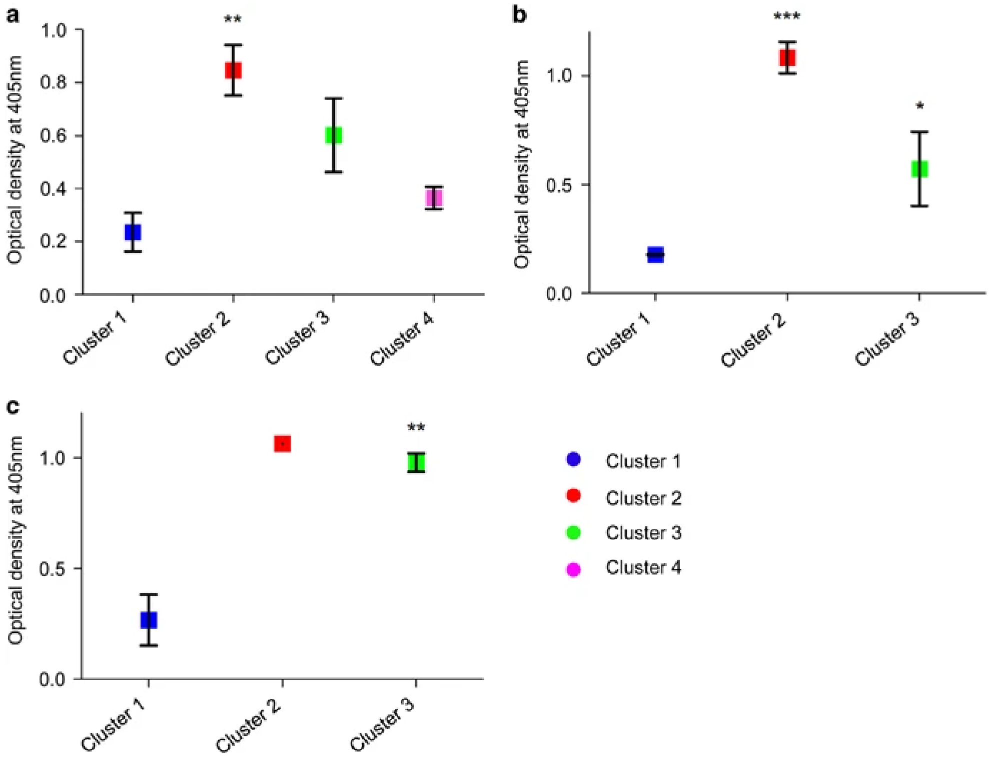

Figure 18. Binding reactivity of scFv antibodies retrieved from selected HCDR3 amino-acid sequences in each cluster using NGS. (a) scFv library 1, (b) scFv library 2 and (c) scFv library 3. ANOVA with Turkey's multiple-comparison test was used to compare cluster 1 with other clusters. In library 3, the P-value was calculated using the Mann–Whitney U-test. *P-value <0.05; **P-value <0.01; ***P-value <0.001. ANOVA, analysis of variance.

Construction of Phage-Displayed scFv Library, Biopanning, Selection of Positive Clones, Next-Generation Sequencing (NGS), And Establishment of TR Data Set

Chickens were immunized with recombinant mouse c-Met-Fc chimera. Spleen, bone marrow, and bursa of Fabricius were harvested from the immunized chickens and total RNA was prepared to generate a phage-displayed scFv library with a complexity of 4.96 × 10^9. Four rounds of bio-panning were performed using antigen-coated magnetic beads. After the final round of bio-panning, the phage pool was subjected to high-throughput clonal retrieval using TR technology. From the TR analysis, 641 clones with unique V_H and V_L pairs were identified. These phage clones were rescued and subjected to phage ELISA. Out of 641 phage clones, 149 clones showed reactivity to c-Met with statistical differences from non-reactive clones (data not shown) designated as AR clones. We used the binding reactivity of the 641 clones as a part of the TR data set.

After arranging the phage-displayed scFv library and each round of bio-panning, phagemid DNA (rounds 0-4) was prepared using bacterial pellets obtained after centrifugation of overnight culture supernatant. From these five sets of phagemid DNA, gene fragments encoding HCDR3 and LCDR3 were amplified and subjected to NGS analysis using the HiSeq platform. After NGS data pre-processing, we defined valid clonotypes as unique CDR3 sequences with read counts of two or higher in any set of phagemid DNA, and we were able to retrieve 860,207 HCDR3 clonotypes and 443,292 LCDR3 clonotypes across the entire bio-panning phase (Table 9). Clonal abundance throughout bio-panning stages was determined by counting the number of times that a clonotype appeared in each bio-panning round. The clonal abundance of clonotypes matching to scFv clones found in TR analysis was used as another part of the TR data set. We also amplified V_H and V_L gene fragments from five sets of phagemid DNA and subjected them to NGS analysis using the MiSeq platform.

Table 9. Number of CDR3 clonotypes obtained from the bio-panning procedure.

Clonotypes	Round 0	Round 1	Round 2	Round 3	Round 4	Total
HCDR3	390,814	395,459	402,854	311,678	308,547	860,207
LCDR3	272,317	253,899	250,630	187,314	117,239	443,292

Establishing Random Forest (RF) Binding Reactivity Prediction Model

We compared random forest, regularized discriminant analysis, linear discriminant analysis, support vector machine, naïve bayes, and AdaBoost classification trees for their accuracy and kappa score distributions. We found out that the random forest algorithm was best suited for binder predictions of HCDR3 clonotypes with the mean accuracy of 89.69% and mean Cohen's kappa value of 0.45 (Table 10 and Table 11). While regularized discriminant analysis did perform marginally better in the LCDR3 clonotypes, random forest showed more potential for improvement with manual tuning when consulting maximum accuracy and Cohen's kappa value (Table 12 and Table 13). With these observations, we decided to adopt random forest models to establish a binding reactivity prediction model (Figure 19).

Utilizing the TR data set, two separate RF models were trained for HCDR3 and LCDR3 clonotypes. The algorithm was instructed to treat the clonal abundance of clonotypes in the five sets of phagemid DNA (round 0–4) as predictor variables and the binding reactivity as the response variable. Thus, each unique clonotype in our TR data set was individually labelled with that clonotype's abundance at each of the bio-panning rounds and its binding reactivity. Before the training of each new RF model, the TR data set was divided into a training data set and a validation data set. After training the RF model using the training set, the validation set was presented to the RF model, and RF model accuracy in predicting clonotype binding reactivity was determined.

To determine the optimum training parameters for our RF model, 7200 RF models were evaluated. Optimizing for sensitivity, the ideal parameters for the HCDR3 RF model were found to be a 75% sampling ratio of the TR data set, mtry of 4, and ntree of 500. The performance metrics of 10 RF models using those parameters were: (1) mean accuracy of 90.48%, (2) mean sensitivity of 44.36%, and (3) mean specificity of 97.61%. Optimizing for accuracy, the ideal parameters for the LCDR3 RF model were found to be a 65% sampling ratio of the TR data set, mtry of 2, and ntree of 500. Once again, the performance metrics of 10 LCDR3 RF models using those parameters were: (1) mean accuracy of 86.47%, (2) sensitivity of 55.98%, and (3) specificity of 94.90%.

Table 10. Accuracy score distributions of random forest (RF), regularized discriminant analysis (RDA), linear discriminant analysis (LDA), support vector machines (SVM), naïve bayes (NB), AdaBoost Classification Trees (ADA) for HCDR3 binding reactivity predictions.

Models	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
RF	0.851	0.868	0.896	0.895	0.918	0.962
RDA	0.830	0.867	0.884	0.882	0.906	0.924
LDA	0.830	0.849	0.858	0.859	0.865	0.907
SVM	0.833	0.849	0.851	0.858	0.865	0.886
NB	0.773	0.830	0.847	0.847	0.886	0.905
ADA	0.132	0.134	0.148	0.143	0.150	0.150

Table 11. Kappa score distributions of RF, RDA, LDA, SVM, NB, and ADA for HCDR3 binding reactivity predictions.

Models	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
RF	0.149	0.367	0.426	0.446	0.564	0.835
RDA	-0.034	0.195	0.306	0.308	0.488	0.629
LDA	-0.034	0.000	0.000	0.040	0.000	0.505
SVM	-0.034	0.000	0.000	0.026	0.000	0.224
NB	0.015	0.133	0.301	0.290	0.434	0.612
ADA	0.000	0.000	0.000	0.000	0.000	0.000

Table 12. Accuracy score distributions of RF, RDA, LDA, SVM, NB, and ADA for LCDR3 binding reactivity predictions.

Models	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
RDA	0.826	0.846	0.876	0.872	0.898	0.924
RF	0.803	0.846	0.849	0.857	0.867	0.943
LDA	0.769	0.788	0.803	0.804	0.826	0.846
NB	0.750	0.776	0.805	0.802	0.825	0.865
SVM	0.769	0.769	0.773	0.774	0.773	0.788
ADA	0.115	0.169	0.180	0.182	0.210	0.230

Table 13. Kappa score distributions of RF, RDA, LDA, SVM, NB, and ADA for LCDR3 binding reactivity predictions.

Models	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
RDA	0.338	0.476	0.596	0.579	0.675	0.755
RF	0.214	0.482	0.532	0.535	0.586	0.833
LDA	0.000	0.122	0.135	0.195	0.259	0.434
NB	0.000	0.156	0.235	0.244	0.340	0.523
SVM	0.000	0.000	0.000	0.000	0.000	0.000
ADA	-0.251	-0.178	-0.141	-0.130	-0.070	0.000

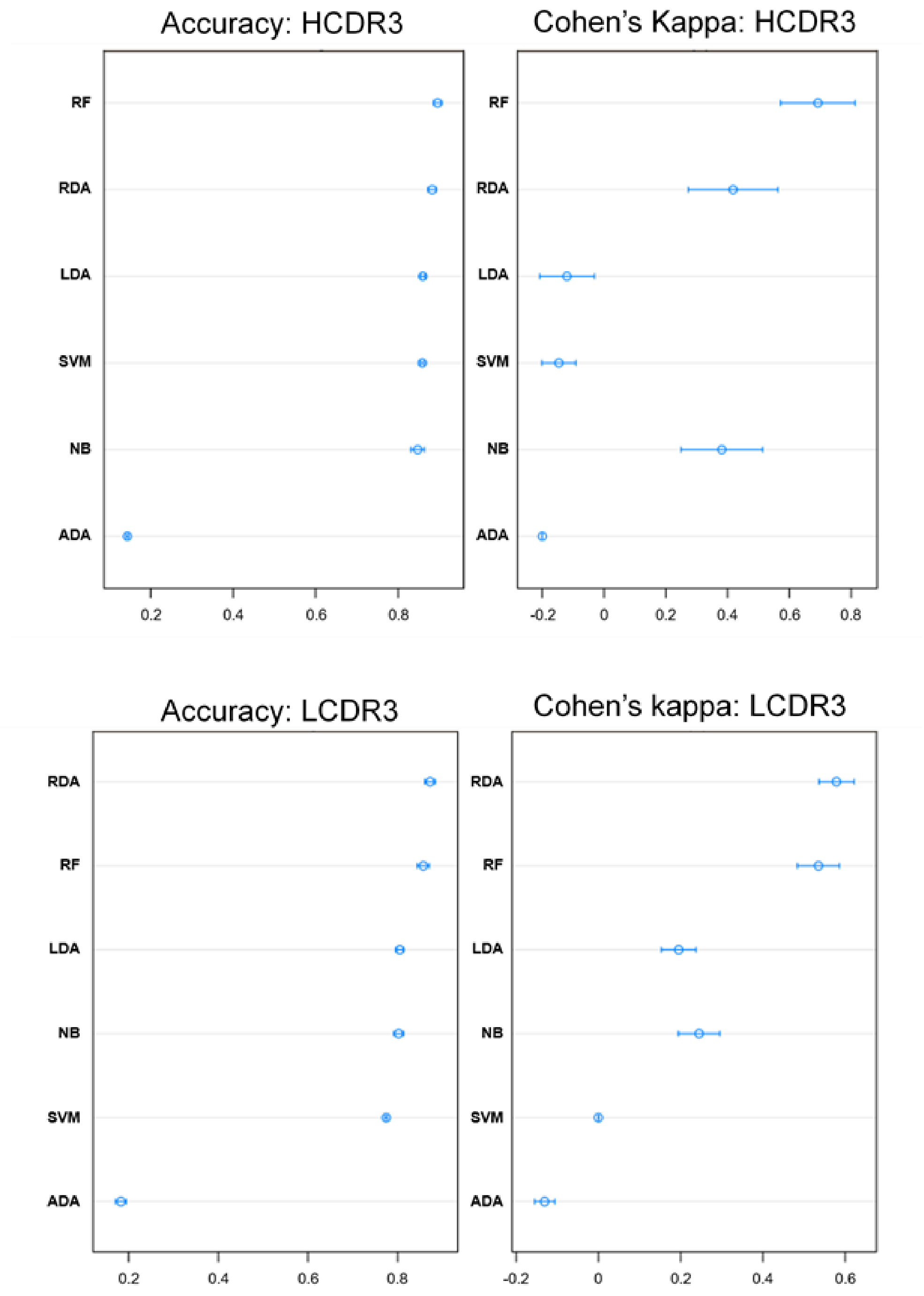

Figure 19. Evaluation of 6 prediction models using training data sets. RF: random forest, RDA: regularized discriminant analysis, LDA: linear discriminant analysis, NB: naïve bayes, SVM: support vector machine, ADA: AdaBoost classification trees. Accuracy and Cohen's kappa value were calculated and plotted as an evaluation metric.

Measurement of the Minimum Depth Value of a Predictor Variable

The minimum depth of a predictor variable can be interpreted as a measure of the variable importance. We extracted the minimum depth value of each predictor variable from the 500 decision trees that compromised our RF model. Of note is that, in our HCDR3 RF model, our predictor variable representing CDR3 clonal abundance in round 3 of bio-panning was most likely to appear at the root node of our decision trees appearing in 360 instances out of our 500 decision trees, and, consequently, had the lowest mean minimum depth of 0.46. In our LCDR3 RF model, our predictor variable representing CDR3 clonal abundance in round 4 of bio-panning was most likely to appear at the root node of our decision trees appearing in 195 instances out of 500 decision trees and, consequently, had the lowest minimum depth of 1.16 (Figure 20). In accordance with these observations, the Shannon entropy (SE) representing clonal diversity dropped at round 3 in the case of HCDR3 clonotypes while the SE of LCDR3 significantly dropped at round 4 (Table 14 and Figure 21).

We also observed the interaction of our predictor variables taking place within the decision trees. Variable interactions are regarded as taking a sub-tree of two nodes and considering it as a single node. We can then look at the minimum depth value of that sub-tree to gauge the interaction's importance in classifying its input. In our HCDR3 RF model, the top four most influential interactions all involved clonal abundance in round 3 of bio-panning as the root node. The most influential interaction took place between round 3 and round 1 with a minimum depth value of 0.84 (Table 15). In our LCDR3 RF model, three of the top four most influential interactions involved clonal abundance in round 4 of bio-panning as the root node. The most influential interaction took place between round 4 and round 0 with a minimum depth value of 1.18 (Table 16). Using the training data set, the clonal abundance of HCDR3 clonotype in round 3 and round 1 and that of LCDR3 clonotype in round 4 and round 0 were plotted in Figure 22a and 22b, which shows significant correlation.

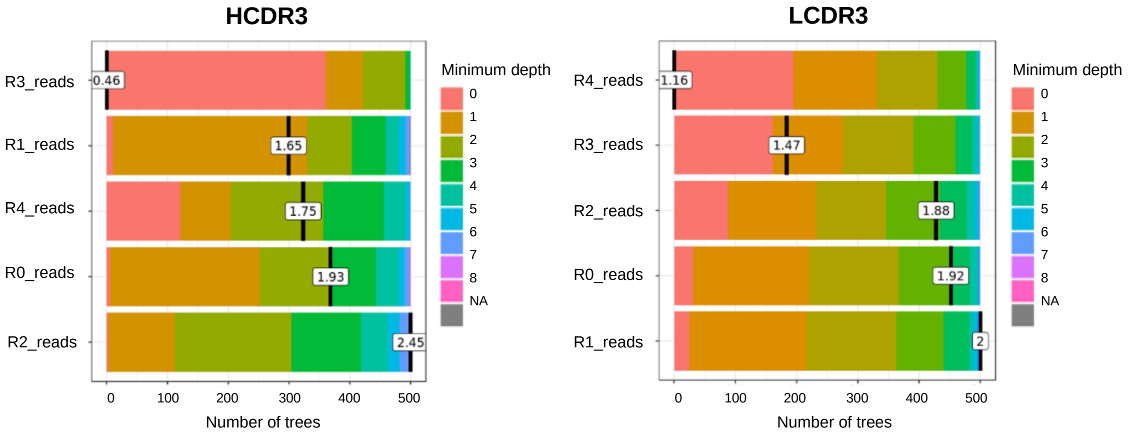

Figure 20. Distribution of the minimum depth of predictor variables (clonal abundance at round 0-4 of bio-panning) from individual decision trees in the RF prediction model for CDR3 clonotypes. Minimum depth value is colored according to its depth and mean value is calculated and displayed at points.

Table 14. Biopanning titer following four rounds of biopanning.

	0.05% PBST wash	antigen: mouse c-Met	
		input titer (PFU/mL)	output titer (PFU/mL)
Round 0	-	4.96×10^{9}	
Round 1	x 1	2.44×10^{11}	6.26×10^{7}
Round 2	x 3	3.28×10^{11}	8.80×10^{4}
Round 3	x 3	2.51×10^{12}	1.12×10^{7}
Round 4	x 5	2.68×10^{11}	1.24×10^{7}

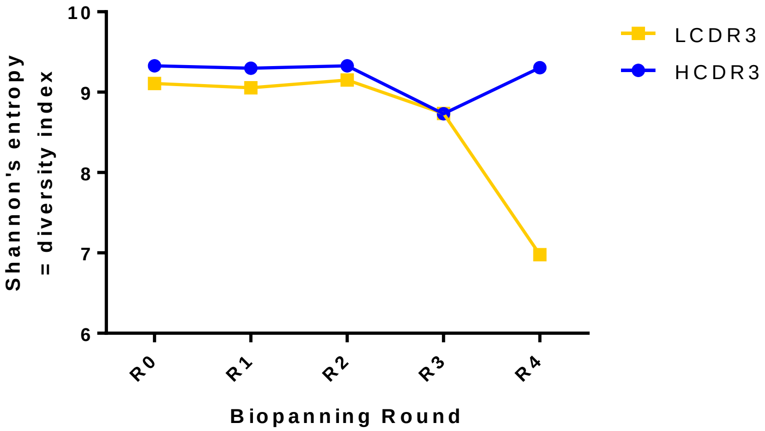

Figure 21. Shannon's entropy (SE) change following biopanning procedure. Yellow and blue dots represent SE of LCDR3 and HCDR3, respectively.

Table 15. Mean-minimal depth of each variables and interaction (HCDR3).

root_variable	variable	mean_min_depth	occurrences	interaction
HCDR3.R3	HCDR3.R1	0.836608	470	HCDR3.R3:HCDR3.R1.
HCDR3.R3	HCDR3.R0	1.232051	460	HCDR3.R3:HCDR3.R0.
HCDR3.R3	HCDR3.R4	1.381857	474	HCDR3.R3:HCDR3.R4.
HCDR3.R3	HCDR3.R2	1.489798	471	HCDR3.R3:HCDR3.R2.
HCDR3.R3	HCDR3.R3	1.554127	462	HCDR3.R3:HCDR3.R3.
HCDR3.R0	HCDR3.R3	1.873671	379	HCDR0.R3:HCDR3.R3.
HCDR3.R4	HCDR3.R4	1.922962	416	HCDR3.R4:HCDR3.R4.
HCDR3.R1	HCDR3.R3	1.952568	399	HCDR3.R1:HCDR3.R3.
HCDR3.R1	HCDR3.R4	2.054730	407	HCDR3.R1:HCDR3.R4.
HCDR3.R4	HCDR3.R3	2.058536	397	HCDR3.R4:HCDR3.R3.
HCDR3.R4	HCDR3.R2	2.191051	403	HCDR3.R4:HCDR3.R2.
HCDR3.R1	HCDR3.R3	2.335815	402	HCDR3.R1:HCDR3.R3.
HCDR3.R0	HCDR3.R4	2.371460	367	HCDR3.R0:HCDR3.R4.
HCDR3.R1	HCDR3.R1	2.412306	388	HCDR3.R1:HCDR3.R1.
HCDR3.R1	HCDR3.R2	2.414817	401	HCDR3.R1:HCDR3.R2.
HCDR3.R0	HCDR3.R2	2.527646	375	HCDR3.R0:HCDR3.R2.
HCDR3.R2	HCDR3.R4	2.534346	296	HCDR3.R2:HCDR3.R4.

Table 16. Mean-minimal depth of each variables and interaction (LCDR3).

root_variable	variable	mean_min_depth	occurrences	interaction
LCDR3.R4	LCDR3.R0	1.180377	444	LCDR3.R4:LCDR3.R0
LCDR3.R4	LCDR3.R2	1.428251	446	LCDR3.R4:LCDR3.R2
LCDR3.R3	LCDR3.R0	1.465202	426	LCDR3.R3:LCDR3.R0
LCDR3.R4	LCDR3.R1	1.486852	432	LCDR3.R4:LCDR3.R1
LCDR3.R4	LCDR3.R3	1.549865	436	LCDR3.R4:LCDR3.R3
LCDR3.R3	LCDR3.R1	1.555740	431	LCDR3.R3:LCDR3.R1
LCDR3.R3	LCDR3.R4	1.632188	427	LCDR3.R3:LCDR3.R4
LCDR3.R4	LCDR3.R4	1.633327	435	LCDR3.R4:LCDR3.R4
LCDR3.R3	LCDR3.R3	1.720592	430	LCDR3.R3:LCDR3.R3
LCDR3.R0	LCDR3.R2	1.785722	428	LCDR3.R3:LCDR3.R2
LCDR3.R1	LCDR3.R4	1.932578	391	LCDR3.R0:LCDR3.R4
LCDR3.R1	LCDR3.R4	1.944619	385	LCDR3.R1:LCDR3.R4
LCDR3.R0	LCDR3.R3	2.245964	377	LCDR3.R1:LCDR3.R3
LCDR3.R2	LCDR3.R3	2.255933	375	LCDR3.R0:LCDR3.R3
LCDR3.R2	LCDR3.R3	2.256614	371	LCDR3.R2:LCDR3.R3
LCDR3.R2	LCDR3.R1	2.276000	370	LCDR3.R2:LCDR3.R1
LCDR3.R2	LCDR3.R4	2.279314	377	LCDR3.R2:LCDR3.R4

Measurement of the Minimum Depth Value of a Predictor Variable

Of the 860,207 HCDR3 clonotypes fed into the RF model, 5,780 clonotypes were predicted to be AR. Of the 443,292 LCDR3 clonotypes, 34,703 clonotypes were predicted to be AR. The confidence value of the RF model for each prediction was also obtained. For HCDR3 and LCDR3, 1.70% (98/5,780) and 0.16% (58/34,703) of clonotypes, respectively, were predicted to be AR with a confidence value of more than 0.9. Meanwhile, 0.56% (4,825/854,427) of HCDR3 clonotypes and 41.14% (168,116/408,589) of LCDR3 clonotypes were predicted to be NR with a confidence value over 0.9. When CDR3 clonotypes were visualized with the most important variable interactions together including a confidence value (Figure 22c), clonotypes with higher confidence values were distributed near the axis of the most important variable akin to the distribution of AR clonotypes in the training data set.

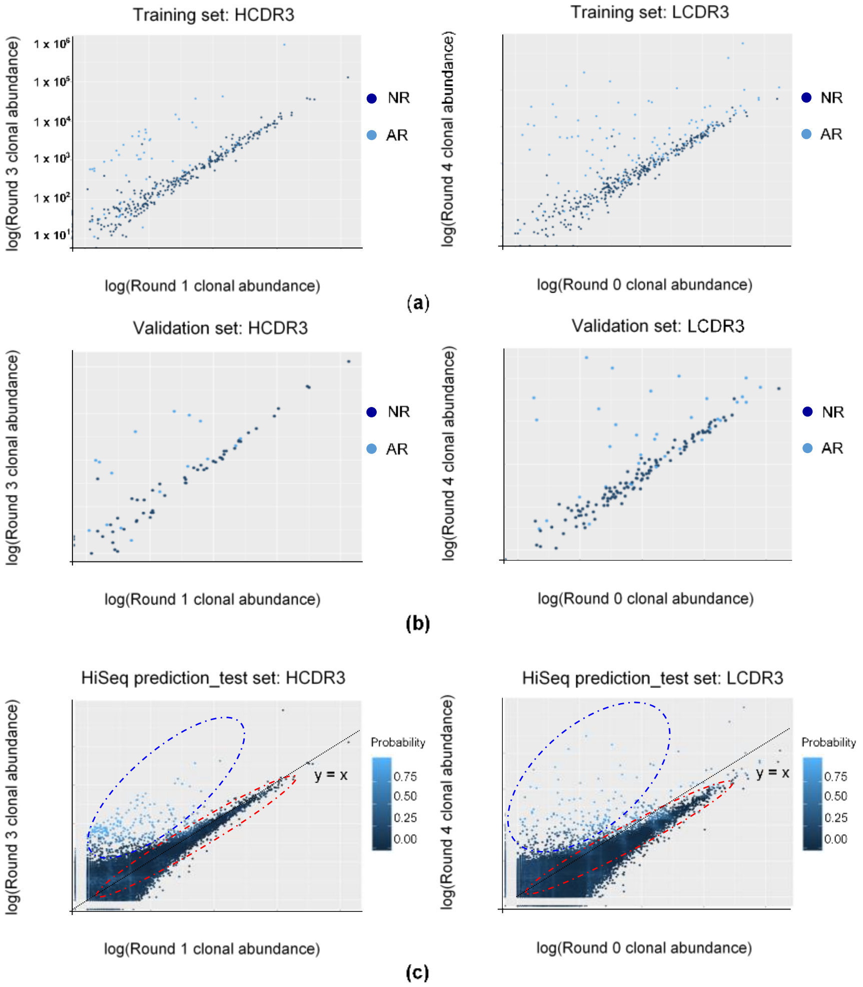

Figure 22. The most influential variable interactions and distributions of CDR3 clonotypes. (a) Clonal abundance at the most influential interaction is plotted with binding property label from training data used in the random forest (RF) prediction model. AR, antigen-reactive, NR, antigen non-reactive. (b) Clonal abundance at the most influential interaction is plotted with a binding property label from validation data used in the RF prediction model. (c) Clonal abundance at the most influential interaction is plotted with confidence value (probability) from HiSeq-identified CDR3 clonotypes. Clonotypes with higher confidence values are distributed near the root variable axis (highlighted with a dashed blue circle) while clonotypes having

lower confidence values are distributed below the y = x axis (dotted line) (highlighted with a dashed red circle).

Antigen Reactivity Validation of In Silico CDR3 Clonotypes in Phage ELISA

We selected 40 HCDR3 AR, 40 LCDR3 AR, 10 HCDR3 NR, and 10 LCDR3 NR clonotypes with the highest confidence values of which whole V_H or V_L gene sequences were available from the NGS data generated from five sets of phagemid DNA using the MiSeq platform (Table 17, 18). After whole V_H and V_L genes were chemically synthesized, V_H and V_L genes of AR clonotypes were used to construct the AR phage-displayed scFv library. In a parallel experiment, the NR phage-displayed scFv library was also constructed using the same scheme. After a single round of bio-panning on antigen-coated magnetic beads, 96 phage clones were randomly selected from the output titer plate of the AR library and subjected to phage ELISA. Fifteen phage clones were found to be AR, which turned out to be 14 scFv clones consisting of five HCDR3 and 11 LCDR3 clonotypes by Sanger sequencing (Figure 23, Table 19). AR5 and AR6 phage clones encoded the same scFv sequence. It was noticeable that three LCDR3 clonotypes were paired with two different HCDR3 clonotypes as in AR1 and AR13, AR2, and AR7, and AR4 and AR14 phage clones showing light chain redundancy. In a parallel experiment, no AR clones were identified from 96 phage clones from the NR library. Sixteen clones were randomly selected and Sanger sequencing was performed to find 13 HCDR3 and nine LCDR3 clonotypes. With these results, we concluded that our RF model can be used to select HCDR3 and LCDR3 AR clonotypes with a significant hit ratio.

We then further validate the RF model by comparing positive rate, clonal diversity and frequency with the conventional colony screening method. As a result, RF model shows higher clonal diversity and positive rate than conventional method either in HCDR3 and LCDR3 (Table 20). Clonal frequency distribution of binder and non-binder clones were compared from each round of bio-panning (Figure 24). At each selection round, frequency distributions were similar between binder and non-binder groups, but frequency ratio of round 1 to 3 showed different pattern in two groups (Figure 25). This result is compatible with the feature importance value observed in RF prediction model. We inferred that RF model generated predicted binders having diverse enrichment pattern, and the result is originated from training data with high clonal frequency variation. However, it is impossible to explain the reason why frequency ratio of round 1 to 3 mostly impacted on the prediction results, which is the intrinsic feature of the supervised learning algorithm.

Table 17. Predicted clones with HCDR3, full variable domain sequences with prediction results and confidence value.

Clone ID	Selected HCDR3	Mapped VH	Prediction	Probability
RFAR1	SAGIGGDCIDA	AVTLDESGGGLQTPGGTLSLVCKASGFTFSSYNMGWVRQAPGKGLEWVAAISNDGSSTGYATAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKSAGIGGDCIDAWGHGTEVIVSS	AR	0.99
RFAR2	CADTGYGCAYCIDA	AVTLDESGGGLQTPGGTLSLVCKASGFTFSSFNMFWVRQAPGKGLEFVASISNTGSYTKYGAAVKGRATISRDDGQSTVRLQLNNLRAEDTGTYYCTRCADTGYGCAYCIDAWGHGTEVIVSS	AR	0.99
RFAR3	TAGTCTTSCNAGAYIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSSFNMFWVRQAPGKGLEFVASISNTGSTTGYGPAVKGRATISRDDGQSTVRLQLNNLRAEDTATYFCAKTAGTCTTSCNAGAYIDAWGHGTEVIVSS	AR	0.96
RFAR4	AVGFACGWCSAGIDA	AVTLDESGGGLQTPGGTLSLVCKASGFSFSSFYMFWVRQAPGKGLEFVAQISSTGSSTDYGSAVKGRATISRDNGQSTLRLQLNNLRAEDTGTYFCAKAVGFACGWCSAGIDAWGHGTEVIVSS	AR	0.96
RFAR5	SADSCATCATYPSEIDT	AVTLDESGGGLQTPGGGLSLVCKASGFTFTDYGMGWMRQAPGKGLEYVAGISNDGSSVAYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCARSADSCATCATYPSEIDTWGHGTEVIVSS	AR	0.95
RFAR6	SGSNWWADSTGNVDA	AVTLDESGGGLQTPGGALSLVCKASGFTFNNYAMNWVRQAPGKGLEYVAAISSSASYTNYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTATYYCAKSGSNWWADSTGNVDAWGHGTEVIVSS	AR	0.95
RFAR7	SPGGYCCAGWIDA	AVTLDESGGGLQTPGGGLSLVCKASGFTFSSYNMGWVRQAPGKGLEWVAGIYSGNRTYYAPAVKGRATISRDNGQSTVRLQLNNLRAEDTATYFCARSPGGYCCAGWIDAWGHGTEVIVSS	AR	0.95
RFAR8	SPGAFTYVSGIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSDYDMAWVRQAPGKGLEFVAGITSDGSNTGYGSAVKGRATISRDNGQSSVRLQLNNLRAEDTGTYICARSPGAFTYVSGIDAWGHGTEVIVSS	AR	0.95
RFAR9	SVTGCGGDYAWCAFGDLDHIDA	AVTLDESGGGLQTPGRALSLVCKASGFTFSSFNMFWVRQAPGKGLEYVAAISSTGSYTKYGAAVQGRATISRDNGQSTVRLQLNNLRAEDTSTYFCAKSVTGCGGDYAWCAFGDLDHIDAWGHGTEVIVSS	AR	0.95
RFAR10	ASGGGYCSWGACIVAWIGT	AVTLDESGGGLQTPGGTLSLVCKASGFSISSYGMGWMRQAPGKGLEFVASISNTGSYTNYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTATYYCAKASGGGYCSWGACIVAWIGTWGHGTEVIVSS	AR	0.95
RFAR11	TTVISCGTLCAGH	AVTLDESGGGLQTPGGTLSLVCKASGFSFSSFYMFWVRQAPGK	AR	0.95

	IDA	GLEFVAQISNTGSSTDYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTAIYYCAKTTVISCGTLCAGHIDAWGHGTEVIVSS		
RFAR12	GASSGSGCAGGLCAGEIDA	AVTLDESGGGLQTPGGTLSLVCKGSGFTFSSVNMGWMRQAPGKGLEWVADINSAGSSTNYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYFCAKGASSGSGCAGGLCAGEIDAWGHGTEVIVSS	AR	0.95
RFAR13	GSGGVDSIDA	AVTLDESGGGLQTPGGAFSLVCKGSGFTFSSFNMFWVRQAPGKGLEYVAGIYYSGSGTGNGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCARGSGGVDSIDAWGHGTEVIVSS	AR	0.95
RFAR14	TADDGNCCGGDNIDA	AVTLDESGGGLQTPGGGLSLVCKASGFTFSDYGMGWVRQAPGKGLEWVAGIYTGSYTGYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKTADDGNCCGGDNIDAWGHGTEVIVSS	AR	0.95
RFAR15	AYSGGFYCAGSLCAAHAGLIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSSYGMFWVRQAPGKGLEWIAGISNSGSYTAYGAVDGRATISRDNGQSTLRLQLNNLRAEDTATYYCAKAYSGGFYCAGSLCAAHAGLIDAWGHGTEVIVSS	AR	0.95

(continued)

Clone ID	Selected HCDR3	Mapped VH	Prediction	Probability
RFAR16	AAASGCAGDNIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSDYGMGWMRQAPGKGLEFVAGIGNTGSWTAYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTATYYCAKAAASGCAGDNIDAWGHGTEVIVSS	AR	0.95
RFAR17	STSDYGGWYGADLDSIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSSFNMFWVRQAPGKGLEWVAQISGDGSTYYAPAVQGRATISRDNGQSTVRLQLNNLRAEDTGTYFCAKSTSDYGGWYGADLDSIDAWGHGTEVIVSS	AR	0.95
RFAR18	TADGGWFGNSAGSIDA	AVTLDESGGGLQTPGGTLSLVCKASGFSISSYTMQWVRQAPGKGLEWVAGISSSGRYTDYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYFCAKTADGGWFGNSAGSIDAWGHGTEVIVSS	AR	0.95
RFAR19	TSGYCGWCGAYNIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSSFNMFWVRQAPGKGLEYVAEISSTGSWTGYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKTSGYCGWCGAYNIDAWGHGTEVIVSS	AR	0.95
RFAR20	SANSGRSASQMDA	AVTLDESEGGLQTPGGALSLVCKASGFTFSDYAMGWVRQAPGKGLEYVASIRGAGSSDTSYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKSANSGRSASQMDAWGHGTEVIVSS	AR	0.95
RFAR21	GGSGYCGWSGYSCVGEIDA	AVTLDESGGGLQTPGGTLSLVCKASGFTFSSSYGMHWVRQAPGKGLEWVAGIYSGGGNTYYAPAVKGRATISRDNGQSTVRLQLNDLRAEDTATYYCTRGGSGYCG	AR	0.95

		WSGYSCVGEIDAWGHGTEVIVSS		
RFAR22	ATGTGYYGSDSYVSSIDA	AVTLDESGGGLQTPGGTLSLVCKGSGFTFSSYDMYWVRQAPGKGLEYVAVISSDGRYTNYGSAVKGRATISKDNGQSTVRLQLNNLRAEDTGTYYCAKATGTGYYGSDSYVSSIDAWGHGTEVIVSS	AR	0.95
RFAR23	SDISWCAWCATDLGQIDA	AVTLDESGGGLQTPGGTLSLVCKASGFTFSSFNMFWVRQAPGKGLEYVASISSADIWTGYGSAVKGRATISRDDGQSTVRLQLNNLRAEDTGTYYCAKSDISWCAWCATDLGQIDAWGHGTEVIVSS	AR	0.95
RFAR24	GAYGHCSGSWCSAGLIDA	AVTLDESGGGLQTPGGTLSLVCKASGFNFSSYQMNWIRQAPGKGLEFVAAINRFGNSTGYAAAVKGRATISRDDGQSTVRLQLNNLRAEDTGTYYCAKGAYGHCSGSWCSAGLIDAWGHGTEVIVSS	AR	0.95
RFAR25	DVYGWCASDCGGSDTIDA	AVTLDESGGGLQTPGGALSLVCKASGFSISSYGMFWVRQAPGKGLEFVAGISSSGRHTDYGSAVKGRATISRDNGQSTMRLQLNNLRAEDTGTYFCAKDVYGWCASDCGGSDTIDAWGHGTEVIVSS	AR	0.95
RFAR26	SAAGYGCTYGSGYGWCVNYIDA	AVTLDESGGGLQTPGRALSLVCKASGFTFSSFNMFWVRQAPGKGLEFVAAISSSGRYTGYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTAIYFCAKSAAGYGCTYGSGYGWCVNYIDAWGHGTEVIVSS	AR	0.95
RFAR27	AAACSGNDCAALLAAGIDA	AVTLDESGGGLQTPGGTLSLVCKASGFTFSSYAMNWVRQAPGKGLEWVGVISDSGNTPKYGPAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKAAACSGNDCAALLAAGIDAWGHGTEVIVSS	AR	0.95
RFAR28	DDSSCIWNTGCTGLIDE	AVTLDESGGGLQTPGGTLSLVCKGSGFTFSSVNMFWVRQAPGKGLEWVAEISTTGRYTNYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKDDSSCIWNTGCTGLIDEWGHGTEVIVSS	AR	0.95
RFAR29	SADGYGWDTAGNMDA	AVTLDESGGGLQTPGGGLSLVCKASGFTFSSNAMGWMRQAPSKGLEFVAAISSSGSGTYYGAAVKGRATISRDDGQSTVRLQLNNLRAEDTAIYFCAKSADGYGWDTAGNMDAWGHGTEVIVSS	AR	0.95
RFAR30	SGTGKYTTGQIDA	AVTLDESGGGLQTPGGTLSLVCKGSGFTFSSFNMFWVRQAPGKGLEYVAEITSGGSYTYYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCARSGTGKYTTGQIDAWGHGTEVIVSS	AR	0.95

(continued)

Clone ID	Selected HCDR3	Mapped VH	Prediction	Probability
RFAR31	TTDSAYCCAGEIDT	AVTLDESGGGLQTPGGTLSLVCKASGFTFSSYGMNWVRQAPGKGLEYVAAISSTGTTTNYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYYCAKTTDSAYCCAGEIDTWGHGTEVIVSS	AR	0.95
RFAR32	TATTCTGCWAGIDSIDA	AVTLDESGGGLQTPGRALSLVCKASGFTFNTYTMFWVRQAPGKGLEFVAGIDNTGSSTGYGPAVQGRATISRDNGQSTVRLQLNNLRAEDTATYYCAKTATTCTGCWAGIDSIDAWGHGTEVIVSS	AR	0.95
RFAR33	SAADYTCGNGGGSCAGSIDA	AVTLDESGGGLQTPGRALSLVCKASGFTFNTYTMFWVRQAPGKGLEWVAQTSNTGRYTAYGPAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYYCAKSAADYTCGNGGGSCAGSIDAWGHGTEVIVSS	AR	0.95
RFAR34	TTGSDYCTLCTGGIDA	AVTLDESGGGLQTPGGGLSLVCKASGFSFSSYDMLWVRQAPGKGLEFVGVISSSGRYTSYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKTTGSDYCTLCTGGIDAWGRGTEVIVSS	AR	0.95
RFAR35	GGGSDSCTACAGSIDA	AVTLDESGGGLQTPGGGLSLICKASGFTFSDYGMGWMRQAPGKGLEYVGVISSSGSTTRYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYYCTRGGGSDSCTACAGSIDAWGHGTEVIVSS	AR	0.95
RFAR36	AAGDSGYAGRIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSSFYMFWVRQAPGKGLEYVAQISGDGSWTYYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYYCAKAAGDSGYAGRIDAWGHGTEVIVSS	AR	0.95
RFAR37	TTCSGSYGWCADSIDA	AVTLDESGGGLQTPGGGLSLVCKASGFTISDYGMGWVRQAPGKGLEYVAQINSAGSYPKYGAAVKGRATISKDNGQSTVRLQLNNLRAEDTATYYCAKTTCSGSYGWCADSIDAWGHGTEVIVSS	AR	0.95
RFAR38	SATTGGAWAGEIDT	AVTLDESGGGLQTPGGGLSLVCKASGFTFSDYQMNWIRQAPGKGLEWVAGISSGGGYTYYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYFCGKSATTGGAWAGEIDTWGHGTEVIVSS	AR	0.95
RFAR39	GCAGCGWSAARIDA	AVTLDESGGGLQTPGGALSLVCKGSGFTFSSYAMFWVRQEPGKGLECVGYINNDGSSTWYATAVKGRATISRDNGQSTVRLQLNNLRAEDTATYYCARGCAGCGWSAARIDAWGHGTEVIVSS	AR	0.95
RFAR40	DTNRDCHSDADSIDA	AVTLDESGGGLQTPGGALSLVCKASGFTFSSYAMNWVRQAPGKGLEWVGGIGSTGSGTYYAPAVQGRATISRDNGQSTVRLQLNNLRAEDTGTYYCAKDTNRDCHSDADSIDAWGHGTEVIVSS	AR	0.95
RFNR1	DAYGYNGWRAGSI	AVTLDESGGGLQTPGGTLSLVCKGSGFTFSSVNMAWVRQAPGKG	NR	0.00

	DA	LEFVAEISSDAGSWTAYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYFCAKDAYGYNGWRAGSIDAWGHGTEVIVSS		
RFNR2	NSGSGGWITDTGRIDA	AVTLDESGGGLQMPGGALSLVCKASGFTFSSYEMQWVRQAPGKGLEWVAGIYSGGTTTSYGPAVKGRATISRDDGQSTVRLQLNNLRAEDTGTYYCAKNSGSGGWITDTGRIDAWGHGTEVIVSS	NR	0.00
RFNR3	SADNGWNTAGRIDA	AVTLDESGGGLQTPGGTLSLICKASGFTFSSVNMGWVRQAPGKGLEFIAQITSRGSSTYYAPAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYYCARSADNGWNTAGRIDAWGHGTEVIVSS	NR	0.00
RFNR4	AAGSGTGWSAGGIDA	AVTLDESGGGLQTPGGALSLVCKGSGFTFNSYAMQWVRQAPGKGLEWVAGISGSGSYTAYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTATYYCAKAAGSGTGWSAGGIDAWGHGTEVIVSS	NR	0.00
RFNR5	SGDAATPDAGGIDA	AVTLDESGGGLQTPGGGLSLVCKGSGFTFSSFNMFWVRQAPGKGLEFVAAINSGGRYTGYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTGIYYCARSGDAATPDAGGIDAWGHGTEVIVSS	NR	0.00
RFNR6	SGYGGYDGSNIDA	AVTLDESGGGLQTPGGGLSLVCKASGFTFSSHGMGWVRQAPGKGLEWVAGIYSGGRYTYYGAAVKGRATISRDNGQSTVRLQLNNLRAEDTAIYYCAKSGYGGYDGSNIDAWGHGTEVIVSS	NR	0.00
RFNR7	ATYAGSGCCDNIDA	AVTLDESGGGLQTPGGVLSLVCKASGFDFSNNDMAWVRQAPGKGLEFVADISSGGGSYTYYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTATYFCARATYAGSGCCDNIDAWGHGTEVIVSS	NR	0.00
RFNR8	GACGGGCYTATFIGTIDV	AVTLDESGGGLQTPGGTLSLVCKGSGFTFSSVNMGWMRQAPGKGLEYVAEISGSGSWTYYAPAVKGRATISRDNGQSTVRLQLNNLRAEDTGTYFCAKGACGGGCYTATFIGTIDVWGHGTEVIVSS	NR	0.00
RFNR9	SAAGYGCAYGWCGDSIDA	AVTLDESGGGLQTPGGALSLVCKASGFSISSYDMAWVRQAPGKGLEFVAGIYSGTTTAYGAAVKGRATISRDDGQSTVRLQLNNLRAEDTATYYCAKSAAGYGCAYGWCGDSIDAWGHGTEVIVSS	NR	0.00
RFNR10	AAGTCYGCSFYATNIDA	AVTLDESGGGLQTPGGALSLVCKGSGFTFSSVNMFWVRQAPGKGLEWVAGIDNTGRYTSYGSAVKGRATISRDNGQSTVRLQLNNLRAEDTAIYFCAKAAGTCYGCSFYATNIDAWGHGTEVIVSS	NR	0.00

Table 18. Predicted clones with LCDR3, full variable domain sequences with prediction results and confidence value.

Clone ID	Selected LCDR3	Mapped V_L	Prediction	Probability
RFAR1	GNYDGSSSVGI	LTQPSSVSANLGGTVKITCSGGSGDYGWYQQKSPGSAPVTVIYWDDERPSGIPS RFSGSTSGSTNTLTITGVQADDEAVYFCGNYDGSSSVGIFGAGTTLTVL	AR	0.99
RFAR2	GSRDSTLAA	LTQPSSVSANLGGTVEITCSGGSGSYGWYQQKSPGSAPVTVIYYNTNRPSDIPS RFSGSKSGSTGTLTITGVQAEDEAVYFCGSRDSTLAAFGAGTTLTVL	AR	0.99
RFAR3	GSYDSSYVGYVGV	LTQPSSVSANLGGTVEITCSGGSGSYGWYQQKSPGSAPVTVIYNDNQRPSNIPS RFSGALSGSTATLTITGVQAEDEAVYYCGSYDSSYVGYVGVFGAGTTLTVL	AR	0.99
RFAR4	GNKDN	LTQPSSVSANPGGTVEITCSGGSGSYGWFQQKAPGSAPVTLIYANTNRPSDIPS RFSGSKSGSTNTLTITGVQADDEAVYYCGNKDNFGAGTTLTVL	AR	0.99
RFAR5	GGYDSTYAGL	LTQPSSVSANLGGTVEITCSGGSYYGWYQQKSPGSAPVTLIYNNDKRPSDIPS RFSGSKSGSTGTLTITGVRAEDEAVYYCGGYDSTYAGLFGAGTTLTVL	AR	0.99
RFAR6	GTADSSGTV	LTQPSSVSANPGETVKITCSGGGSSSYYGWYQQKSPGSAPVTLIYESNKRPSDIPS RFSGSKSGSTATLTITGVQADDEAVYYCGTADSSGTVFGAGTTLTVL	AR	0.99
RFAR7	GSRDSSYVPI	LTQPSSVSANLGGTVEITCSGGSGSYGWYQQKSPGSAPVTVIYYNTNRPSDIPS RFSGSKSGSTHTLTITGVRAEDEAVYFCGSRDSSYVPIFGAGTTLTVL	AR	0.99
RFAR8	GSWDSSSEGDSGYAGI	LTQPSSVSANPGETVKITCSGSRNSYGWYQQKSPGSAPVTVIYWNSNRPSGIPS RFSGSTSGSTGTLTITGVQADDEAVYYCGSWDSSSEGDSGYAGIFGAGTTLTVL	AR	0.99
RFAR9	GAYDSSYIGI	LTQPSSVSANLGGTVKITCSGGSSGYGWYQQKSPGSAPVTVIYSNTNRPSDIPS RFSGSKSGSTGTLTITGVQAEDEAVYYCGAYDSSYIGIFGAGTTLTVL	AR	0.99
RFAR10	GSFDSSYVGM	LTQPSSVSANPGETVKITCSGGSGNYGWYQQKSPGSAPVTVIYDSSSRPSDIPS RFSGSTSGSTSTLTITGVQADDEAVYYCGSFDSSYVGMFGAGTTLTVL	AR	0.99
RFAR11	GSIDSNYDGI	LTQPSSVSANPGETVKLICSGSSGDYGWYQQKSPGSAPVTVIYDNTNRPSNIPS RFSGSLSGSTNTLSITGVQVEDEAVYFCGSIDSNYDGIFGAGTTLTVL	AR	0.99
RFAR12	GSRDNSSAST	LTQPSSVSANPGETVEITCSGSSSGYGYGWYQQKSPGSAPVTLIYSNDKRPSDIPS RFSGSKSGSTGTLTITGVRAEDEAVYFCGSRDNSSASTFGAGTTLTVL	AR	0.99
RFAR13	GSFDSSSDSGYVGI	LTQPSSVSANPGETVKITCSGGSNNYGWYQQKSPGSAPVTVIYDNTNRPSDIPS RFSGSASGSASTLTITGVQADDEAVYYCGSFDSSSDSGYVGIFGAGTALTVL	AR	0.99
RFAR14	GSYDSSYVGL	LTQPSSVSANLGGTVEITCSGGSSNEYGWYQQKAPGSAPVTLIYDNTNRPSDIPS RFSGSKSGSTGTLTIAGVQAEDEAVYFCGSYDSSYVGLFGAGTTLTVL	AR	0.98
RFAR15	GSTDSSNTDI	LTQPSSVSAKPGGTVEITCSGGSGSYGWFQQKSPGSAPVTLIYANTNRPSDIPS RFSGSKSGSTATLTITGVQAEDEAIYYCGSTDSSNTDIFGAGTTLTVL	AR	0.98

(continued)

Clone	Selected	Mapped V_L	Prediction	Probability
ID	LCDR3			
RFAR16	GSRAGSSI	LTQPSSVSANPGETVKITCSGSSGSYGWYQQKSPGSAPVTVIYYNDKRPSDIPSRFSGSKSGSTGTLTITGVQAEDEAVYFCGSRAGSSIFGAGTTLTVL	AR	0.98
RFAR17	GSYDSSYDGV	LTQPSSVSANPGETVKITCSGSSGYGYGWYQQKSPGSAPVTVIYYNDKRPSNIPSRFSGSKSGSTATLTITGVRADDEAVYFCGSYDSSYDGVFGAGTTLTVL	AR	0.98
RFAR18	GNGDRSSTTGI	LTQPSSVSANLGETVKITCSGGSGSYGWFQQKSPGSAPVTVIYSNDKRPSDIPSRFSGSKSGSTGTLTITGVQADDEAVYYCGNGDRSSTTGIFGAGTTLTVL	AR	0.97
RFAR19	GNEDISGI	LTQPSSVSANPGETVKITCSGGSYKYGWFQQKSPGSAPVTVIYYNDKRPSNIPSRFSGSKSGSTATLTITGVQADDEAVYYCGNEDISGIFGAGTSLTVL	AR	0.97
RFAR20	GSFDSSYTGI	LTQPSSVSANLGGTVKITCSGSSGSYGYGWYQQKSPGSAPVTVIYSNNQRPSNIPSRFSGSTSGSTGTLTITGVRAEDEAVYYCGSFDSSYTGIFGAGTTLTVL	AR	0.97
RFAR21	GSTDSSRTDT	LTQPSSVSANLGGTVKITCSGSSGSYGWYQQKSPGSAPVTLIYQNTKRPSDIPSRFSGSKSGSTGTLTITGVQAEDEAVYYCGSTDSSRTDTFGAGTTLTVL	AR	0.97
RFAR22	GSIDSRYVGI	LTQPSSVSANLGETVKITCSGGSYSYGWYQQKAPGSAPVTLIYDNTNRPSDIPSRFSGSKSGSTHTLTITGVQADDEAVYFCGSIDSRYVGIFGAGTTLTVL	AR	0.97
RFAR23	GGYDGSSAA	LTQPSSVSANPGGTVEITCSGGSGNNYGWFQQKSPGSTPVTVIYNNDKRPSDIPSRFSGSKSGSTATLTITGVQADDEAVYYCGGYDGSSAAFGAGTTLTVL	AR	0.97
RFAR24	ANYDSSTDI	LTQPSSVSANPGETVKITCSGGSSGYGYGWFQQKSPGSAPVTLIYYNDKRPSDIPSRFSGSTSGSTSTLTITGVQADDEAVYYCANYDSSTDIFGAGTTLTVL	AR	0.97
RFAR25	GSYDSTYAGM	LTQPSSVSANPGETVKITCSGGSGSYGWYQQKSPGSAPVTVIYYNYKRPSDIPSRFSGSASGSTATLTITGVQAEDEAVYYCGSYDSTYAGMFGAGTTLTVL	AR	0.97
RFAR26	GSGDSSGTEAA	LTQPSSVSANPGGTVEITCSGSSGSYGWYQQKSPGSAPVTLIYANTNRPSNIPSRFSGSTSGSTATLTITGVQADDEAVYYCGSGDSSGTEAAFGAGTTLTVL	AR	0.97
RFAR27	GSEDSSGAGYVGI	LTQPSSVSANPGETVKITCSGGSYGYSWHQQKSPGSAPVTVIYSSNQRPSDIPSRFSGSTSGSTATLTITGVQADDEAVYFCGSEDSSGAGYVGIFGAGTTLTVL	AR	0.96
RFAR28	GGFDSTDSGYAGI	LTQPSSVSANPGETVKITCSGSTSTYYGWYQQKSPGSAPVTLIYNNNNRPSDIPSRFSGSTSGSTNTLTITGVRAEDEAVYYCGGFDSTDSGYAGIFGAGTTLTVL	AR	0.96
RFAR29	GSADTKYVGI	LTQPSSVSANPGETVEITCSGDSSYYGWYQQKSPGSAPVTVIYDNTNRPSDIPSRFSGSLSGSTNTLTITGVQVEDEAIYFCGSADTKYVGIFGAGTTLTVL	AR	0.96
RFAR30	GSRDSSYLDSGI	LTQPSSVSANLGGTVKITCSGGGSYYGWYQQKAPGSAPVTLIYWNDNRPSDIPSRFSGSKSGSTATLTITGVQADDEAVYYCGSRDSSYLDSGIFGAGTTLTVL	AR	0.95

(continued)

Clone ID	Selected LCDR3	Mapped V_L	Prediction	Probability
RFAR31	GTWDSNTYA	LTQPSSVSANLGETVKITCSGGSGNYGWFQQKAPGSAPVTVIYYDDERPSNIPS RFSGSTSGSTSTLTITGVQVEDEAVYFCGTWDSNTYAGIFGAGTTLTVL	AR	0.95
RFAR32	GSYEDSSYVGI	LTQPSSVSANLGGTVKITCSGGSGSYGWFQQKSPGSVPVTVIYDSSSRPSDIPS RFSGSKSGSTGTLTITGVQAEDEAVYFCGSYEDSSYVGIFGAGTTLTVL	AR	0.95
RFAR33	GSYVSGKYDGI	LTQPSSVSANPGETAKITCSGGYRSYGWYQQKSPGSAPVTLIYSNNQRPSSIPS RFSGSVSVFTHTLTITGVQAEDEAVYYCGSYVSGKYDGIFGAGTTLTVL	AR	0.95
RFAR34	GTADSSTEAI	LTQPSSVSANPGETVKITCSGGSGRYGWFQQKSPGSAPVTVIYWDDERPSNIPS RFSGSTSGSTNTLTITGVQVEDEAVYFCGTADSSTEAIFGAGTTLTVL	AR	0.95
RFAR35	GSYDNTYAGI	LTQPSSVSANLGGTVEITCSGGSGSYGWYQQKAPGSAPVTVIYANTNRPSNIPS RFSGSKSGSTNTLTITGVQAEDEAVYFCGSYDNTYAGIFGAGTTLTVL	AR	0.95
RFAR36	GGYDSSSSSAV	LTQPSSVSANLGGTVKITCSGSSSNNYGWYQQKSPGSTPLTLIYWNDKRPSDIPS RFSGSTSGSTATLTITGVQAEDEAVYFCGGYDSSSSSAVFGAGTTLTVL	AR	0.95
RFAR37	GSYEDSNY	LTQPSSVSANPGETVEITCSGSRTGYGWFQQKSPGSAPVTLIYGSNKRPSNIPS RFSGSKSGSTSTLTITGVQAEDEAVYFCGSYEDSNYFGAGTTLTVL	AR	0.95
RFAR38	GSFDSSYSGI	LTQPSSVSANLGGTVKITCSGGSSGYYGWYQQKSPGSAPVTLIYSNNQRPSNIPS RFSGSGSGSTGTLTITGVRAEDEAVYFCGSFDSSYSGIFGAGTTLTVL	AR	0.95
RFAR39	GDWDSNI	LTQPSSVSANPGETVEITCSGDSNYYGWYQQKAPGSAPVTLIYANTNRPSNIPS RFSGSGSGSTNTLTITGVQAEDEAVYYCGDWDSNIFGAGTTLTVL	AR	0.95
RFAR40	GGYDSSSGA	LTQPSSVSANPGETVKITCSGGGSSRYYGWYQQKAPGSAPVTLIYDNTNRPSNIPS RFSGSKSGSTATLTITGVQAEDEAVYFCGGYDSSSGAFGAGTTLTVL	AR	0.94
RFNR1	GGYDGSTDAGI	LTQPSSVSANPGETVKITCSGSSSSYYGWYQQKSPGSAPVTLIYDNTNRPSDIPS RFSGSKSGSTATLTITGVQADDEAVYFCGGYDGSTDAGIFGAGTTLTVL	NR	0.00
RFNR2	GSTDSSYTDSL	LTQPSSVSANPGETVKITCSGGGSYDYGWYQQKSPGSAPVTVIYNNNKRPSDIPS RFSGALSGSTATLTITGVQADDEAVYFCGSTDSSYTDSLFGAGTTLTVL	NR	0.00
RFNR3	GNEDSSYAGI	LTQPSSVSANLGGTVEITCSGGSGSYGWFQQKAPGSAPVTLIYANTNRPSDIPS RFSGSKSGSTATLIITGVQAEDEAVYFCGNEDSSYAGIFGAGTTLTVL	NR	0.00
RFNR4	GNYADSSST	LTQPSSVSANPGETVKITCSGGTYNYGWYQQKSPGSAPVTVIYDNNKRPSDIPS RFSGALSGSTATLTITGVQADDEAVYFCGNYADSSSTFGAGTTLTVL	NR	0.00
RFNR5	GSADSSSAGI	LTQPSSVSANLGGTVKITCSGSSDSYGWYQQKSPGSAPVTLIYESNKRPSDIPS RFSGSKSGSTGTLTITGVQAEDEAVYYCGSADSSSAGIFGAGTTLTVL	NR	0.00
RFNR6	GSADSSGSGI	LTQPSSVSANPGETVKITCSGGGSYGWYQQKSPSSAPVTLIYTNTNRPSNIPS RFSGSKSGSTGTLTITGVQAEDEAVYFCGSADSSGSGIFGAGTTLTVL	NR	0.00
RFNR7	GSRDSSNVGI	LTQPSSVSANLGGTVEITCSGGGSYGWYQQKSPGSAPVTVIYWNDKRPSDIPS RFSGSKSGSTGTLTITGVQAEDEAVYFCGSRDSSNVGIFGAGTTLTVL	NR	0.00
RFNR8	GSYEGSSGIV	LTQPSSVSANPGETVKITCSGSSGSYGWYQQKSPGSAPVTVIYSNDKRPSDIPS RFSGSASGSTATLTITGVQADDEAVYYCGSYEGSSGIVFGAGTTLTVL	NR	0.00
RFNR9	GSRDSTDSLYVGI	LTQPSSVSANPGETVKITCSGGSSYYAWYQQKSPGSAPVTVIYYNDKRPSDIPS RFSGSTSGSTSTLTITGVQADDEAVYFCGSRDSTDSLYVGIFGAGTTLTVL	NR	0.00
RFNR10	GSADSSTDSGI	LTQPSSVSANPGGTVEITCSGGSSNYGWFQQKAPGSAPVTVIYNNNKRPSDIPS RFSGSKSGSTGTLTITGVQADDEAVYFCGSADSSTDSGIFGAGTTLTVL	NR	0.00

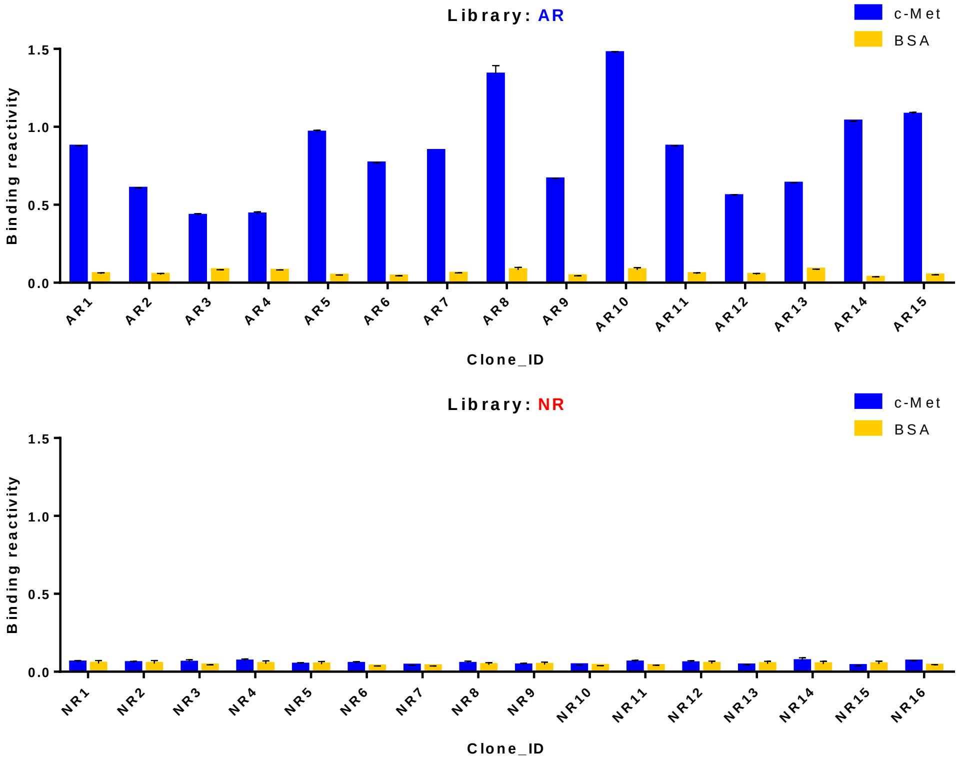

Figure 23. Reactivity of phage-displayed scFv clones in phage ELISA. Binding reactivity of 15 unique clones identified from the AR library and 16 unique clones from the NR library are shown. Wells in microtiter plates were either coated with recombinant mouse c-Met or just blocked with 3% BSA in PBS. Phage clones, HRP-conjugated anti-M13 antibody, and HRP substrate solution were added sequentially with intermittent washing.

Table 19. Amino acid sequences of AR CDR3 clonotypes identified from AR library.

Clone ID	HCDR3 AA* Sequence	LCDR3 AA* Sequence
AR1	GSGGVDSIDA	GSYDNTYAGI
AR2	SADGYGWDTAGNMDA	GSIDSNYDGI
AR3	TAGTCTTSCNAGAYIDA	GGYDGSSAA
AR4	TTCSGSYGWCADSIDA	GAYDSSYIGI
AR5	SADSCATCATYPSEIDT	GSFDSSYVGM
AR6	SADSCATCATYPSEIDT	GSFDSSYVGM
AR7	SADSCATCATYPSEIDT	GSIDSNYDGI
AR8	SADSCATCATYPSEIDT	GSYDSSYVGL
AR9	SADSCATCATYPSEIDT	GSYDSSYDGV
AR10	SADSCATCATYPSEIDT	GSFDSSYTGI
AR11	SADSCATCATYPSEIDT	GSIDSRYVGI
AR12	SADSCATCATYPSEIDT	GSYDSSYVGYVGV
AR13	SADSCATCATYPSEIDT	GSYDNTYAGI
AR14	SADSCATCATYPSEIDT	GGYDSSSGA
AR15	SADSCATCATYPSEIDT	GAYDSSYIGI

* AA: amino acid.

Table 20. Validation of positive rate and clonal diversity from conventional colony screening method and RF prediction model.

Library	Region	Positive rate	Species richness (normalized)
Training	scFv	149/641 (23.24%)	4.88
	VH	91/582 (15.64%)	3.28
	HCDR3	80/531 (15.07%)	3.16
	VL	145/634 (22.87%)	4.80
	LCDR3	123/524 (23.47%)	4.93
RF prediction model	scFv	330/376 (87.77%)	18.44
	VH, HCDR3	15/21 (71.43%)	15
	VL, LCDR3	22/28 (78.57%)	16.54

scFv, single chain variable fragment; VH, variable heavy chain; HCDR3, heavy chain complementary region 3; VL, variable light chain; LCDR3, light chain complementary region 3.

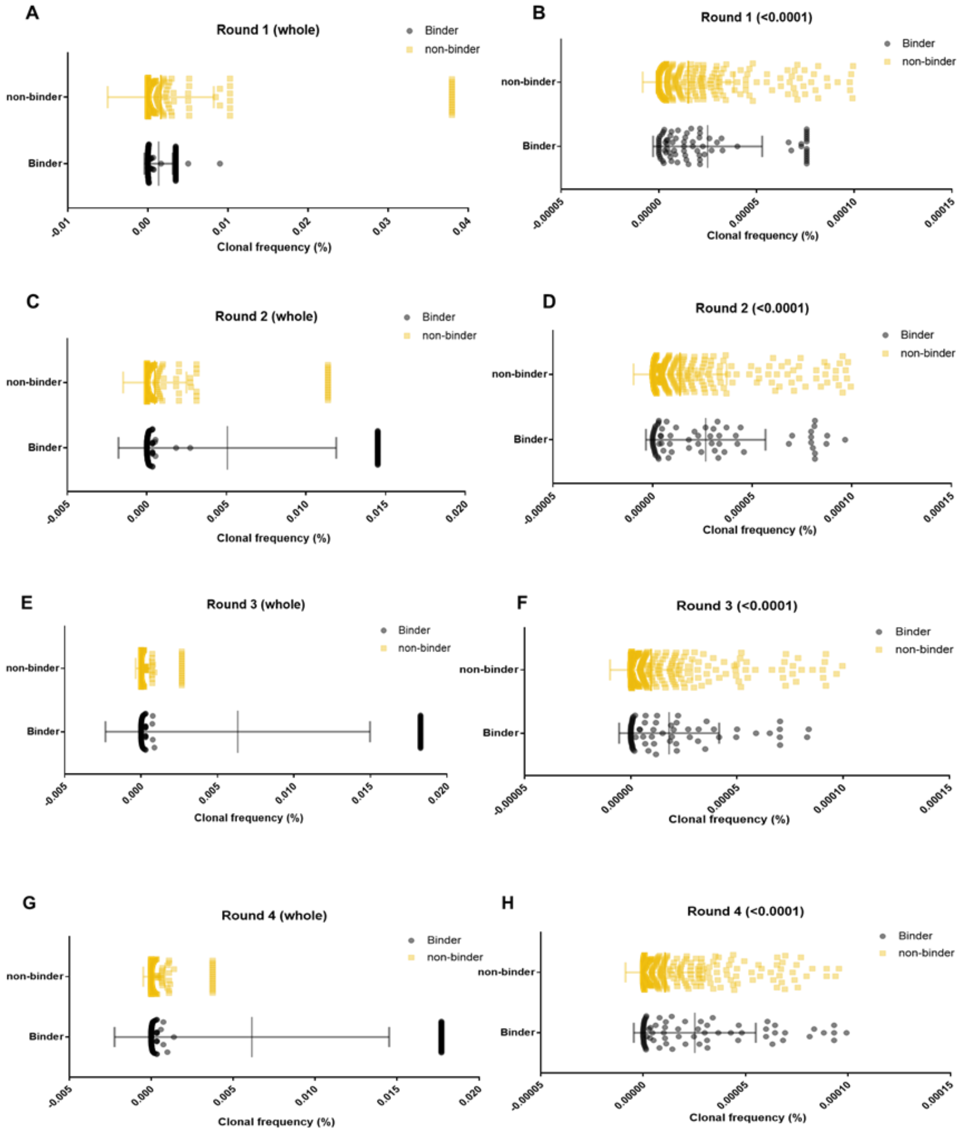

Figure 24. Clonal frequency distribution in the training data. Clonal frequency of each clonotype were calculated and plotted from entire clonotypes (A, C, E, G) and rare clonotypes (clonotype having clonal frequency less than 0.0001) (B, D, F, G). Yellow dots and black dots represent the non-binder and binder clones respectively.

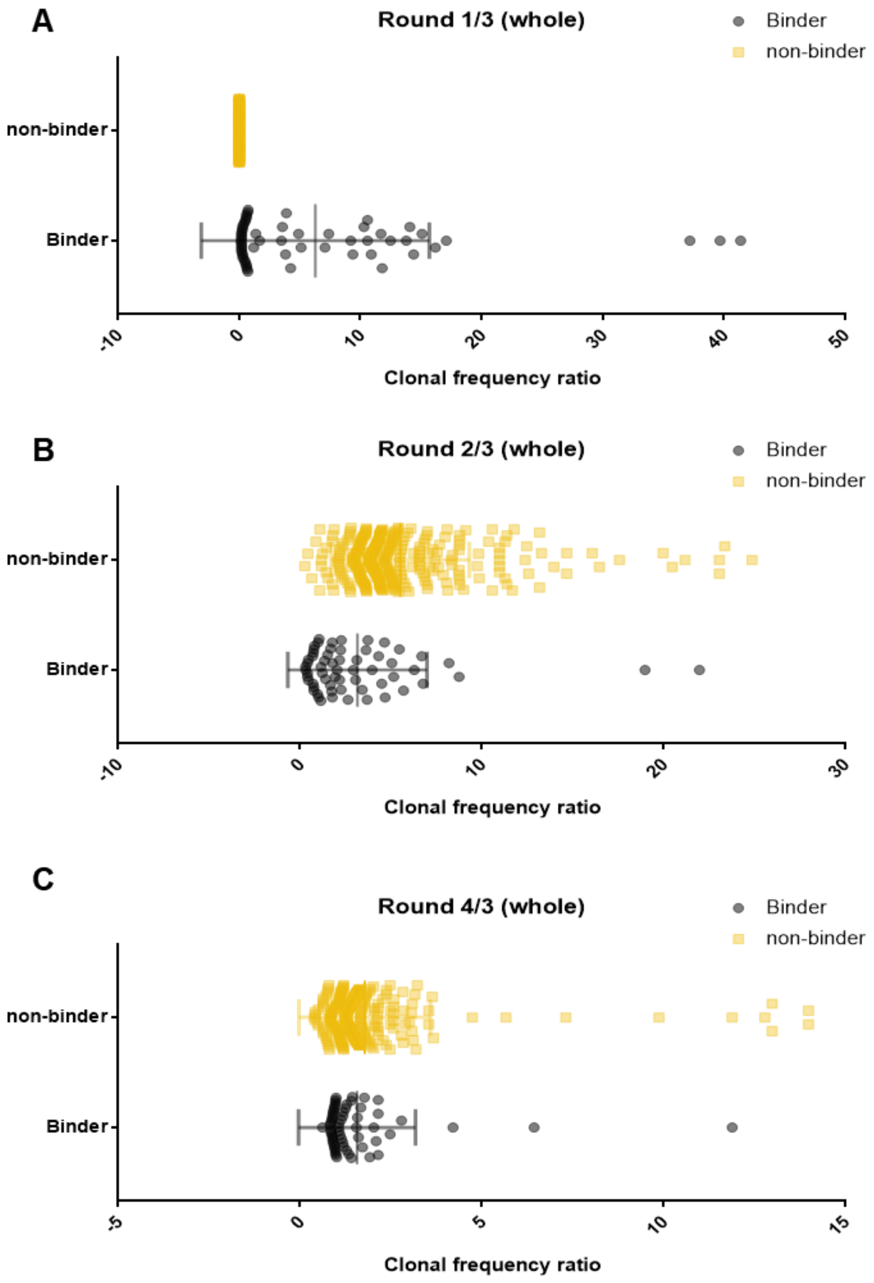

Figure 25. Distribution of clonal frequency ratio in the training data. Clonal frequency ratio of round 1 (A), 2 (B), 4 (C) to round 3 was calculated and plotted. Yellow dots and black dots represent the non-binder and binder clones respectively.

4.4. Discussion

Despite the promise of targeting the HGF-c-Met signaling pathway for cancer therapy, no specific therapeutic agent has been approved for clinical use. Small molecule inhibitors specific to c-Met are yet to be approved, and only nonspecific tyrosine kinase inhibitors inhibiting c-Met are available (Table 21)[130]. Recombinant protein (truncated HGF, decoy c-Met) was not successful in clinical trials due to several factors, including short half-life and low target affinity limiting the intended efficacy[131]. Several HGF-neutralizing antibodies have been developed with two currently active in clinical trials[132]. However, the inhibitory targeting of c-Met by an antibody has been difficult since the bivalency of antibodies often induces receptor dimerization, which potentially causes cancer cell proliferation and migration. As such, both a monovalent form of antibody blocking its interaction with HGF and a bivalent antibody inducing receptor internalization have been developed and tested in clinical trials unsuccessfully[126,133]. Recently, an anti-EGFR x c-Met bispecific antibody monovalent to each target came under clinical development, which should inhibit the ligand interaction and induce the internalization of both receptors[134,135]. Besides blocking the interaction with ligand and receptor internalization, other mechanisms of actions for therapeutic antibody binding to targets on cancer cells were also reported, which include complement-dependent cell cytotoxicity as observed in rituximab[136], antibody-dependent cell cytotoxicity seen with obinutuzumab[137], and phagocytosis of antibody-opsonized tumor cells[138]. Antibodies are also used to deliver cytotoxic payloads into cancer cells such as with T-DM1[139], and cross-linking cancer cells to cytotoxic T cells with blinatumomab[140]. Furthermore, antibodies are used as a cancer cell-targeting component in chimeric antigen receptor T cell therapy, as seen with tisagenlecleucel and axicabtagene ciloleucel[27]. Additionally, it is well known that the antibody epitope and binding characteristics critically influence efficacy for all these various modes of action[141,142]. Therefore, it is crucial to develop a significant number of antibodies to a target and characterize their performance. However, antibody selection technologies, including conventional hybridoma and display technologies such as phage, ribosomal, and bacterial, all have their own limitations regarding high-throughput capabilities.

After George P. Smith and Gregory P. Winter successfully displayed recombinant peptides and antibodies at the pIII protein of the M13 phage[143], this powerful technology has evolved and been actively applied toward therapeutic antibody discovery[144,145]. Currently, over 80 antibodies derived from phage display libraries have entered clinical studies with 10 of these granted marketing authorization[146]. Since Ravn U et al. demonstrated the potential for NGS analysis in the phage-displayed antibody repertoire in 2010, numerous groups have leveraged similar strategies for discovering antibodies reactive to specific antigens[64,147-155]. The next hurdle to overcome after the identification of in silico antibody sequences in NGS data was the low-throughput nature of chemically synthesizing all antibody sequences and individually testing their reactivity. Recently, we introduced a method for combining NGS analysis and individual antibody sequence identification with the isolation of their physical DNA, which was named TR technology[63]. To reduce the burden of expressing all of the antibodies, we also devised a way of predicting antigen reactivity toward antigens by clustering antibody clonotypes with their patterns of enrichment or restriction through bio-panning rounds, and then combining TR with clustering and testing reactivity for a significant number of clones.

Table 21. Clinical usage of small molecule inhibitors targeting c-Met in cancer therapy.

Drug Name	Targets	FDA Approval Status	Approved Year
Tivantinib	c-Met, microtubule	None	N.A.*
Foretinib	c-Met, VEGFR-2*	None	N.A.
Cabozantinib	c-Met, VEGFR, Axl	Medullary thyroid cancer Advanced renal cell carcinoma Hepatocellular carcinoma	2012 2016 2019
Crizotinib	c-Met, ALK*, ROS1, RON*	ALK or ROS-1 positive NSCLC*	2011
Capmatinib	c-Met, EGFR*, ErbB-3	None	N.A.
AMG337	c-Met	None	N.A.
AZD6094	c-Met	None	N.A.
BMS777607/ ASLAN002	c-Met, Axl, Tyro3, RON	None	N.A.
Glesatinib	c-Met, Axl	None	N.A.
Tepotinib	c-Met	None	N.A.

* VEGFR-2: Vascular endothelial growth factor-2, ALK: Anaplastic lymphoma kinase, RON: Receptor d'Origine nantais, EGFR: Epidermal growth factor receptor, NSCLC: Non-small cell lung cancer, N.A.: not available.

Using these tools and procedures, we believed that it was possible to train a machine learning algorithm to derive in silico AR clonotypes from a repertoire of NGS sequences. To demonstrate this, we performed an in-depth analysis of our bio-panning library with the guidance of our supervised machine learning algorithm trained with large amounts of data sets generated from a high-throughput clone retrieval platform and independent NGS analysis. The RF model utilized is composed of numerous unique decision trees that work together to classify inputs. Each decision tree in an RF model is generated using a bootstrapped sample of the training data and a randomized subset of variables evaluated for the best split at each node of that decision tree. As a result, each RF model decision tree is uniquely generated and makes the model more robust to overfitting compared to other linear classifiers or decision trees. Compared to the more complicated black boxes of artificial neural networks, RF models frequently show similar levels of predictive performance while remaining observable and transparent. By inspecting the composition of decision trees in the RF model, we can extract important measures of input variables to better understand the decision-making process of the algorithm. Our extraction of variable importance measures helped explain the logical processes of our RF prediction model, which consists of complex, randomized interactions of predictor variables and response variables. From these results, we can infer that AR HCDRs are mostly selected in enrichment rounds, while LCDR3s are significantly enriched with selected HCDR3s after additional selective pressure occurs. We can then infer that enrichment of scFv molecules depends on individual chains in different stages of the bio-panning process (first V_H is then significantly biased by V_L). We believe our prediction model may be enhanced to better predict binding reactivity with multiple (high, mid, low) rather than binary (reactive/non-reactive) classifications. It is highly likely that this model can be applied to other display platforms that use bio-panning as the selection process, such as yeast display library for fluorescence-activated cell sorting screening[156]. Recently, artificial intelligence has been applied to predict the physicochemical properties of antibody sequences[157-161] and/or optimize them[162-164].

In summary, we report that machine learning algorithm can provide a way to identify AR antibody clones with a significant hit ratio, which will allow us to better characterize diverse antibodies in greater numbers currently unattainable by traditional methods.

4.5. Methods

Library construction and bio-panning

Three white leghorn chickens were immunized and boosted four times with recombinant human PSA (Fitzgerald, Acton, MA, USA). After the final booster injection, total RNA was extracted from the spleen, bone marrow, and bursa of Fabricius using the TRI Reagent (Invitrogen, Grand Island, NY, USA). First-strand cDNA was synthesized using SuperScript reverse transcriptase with oligo (dT) priming (Invitrogen). Using this cDNA, three phage-displayed libraries of chicken scFvs were constructed using the pComb3XSS phagemid vector[165]. Four rounds of bio-panning were performed to screen scFv clones from the library following a previously reported procedure[165]. For each round of bio-panning, 5×10^6 magnetic beads (Dynabeads M-270 epoxy) (Invitrogen) coated with 1.5 μg recombinant PSA protein were used.

Phage enzyme immunoassay

The scFv-displaying phages were rescued from titer plates after transformation and subjected to phage enzyme immunoassay as described previously[165]. The microtiter plates (Corning, NY, USA) were coated overnight at 4 °C with 20 μL recombinant human Fc-tagged PSA (5 μl ml−1) dissolved in phosphate-buffered saline (PBS). After blocking with 3% bovine serum albumin dissolved in PBS (w/v, PBS-B), the plates were then sequentially incubated with scFv-displaying phages in the culture supernatant, horseradish peroxidase (HRP)-conjugated mouse anti-M13 monoclonal antibody (GE Healthcare, Pittsburg, PA, USA) in PBS-B, and then finally with 2,2′-Azinobis [3-ethylbenzothiazoline-6-sulfonic acid]-diammonium salt (ABTS) substrate solutions (Amresco LLC, Solon, OH, USA), with intermittent washing using 0.05% Tween-20 in PBS (PBST). After incubating the plates at 37 °C for 10 min, the optical density was measured at 405 nm using a microtiter plate reader (Labsystems AiG SL, Barcelona, Spain).

Sanger sequencing analysis

Phagemid DNA from selected clones identified by phage enzyme immunoassays was prepared with a small-scale plasmid preparation kit (Qiagen, Hilden, Germany). The OmpSeq primer (5′-AAGACAGCTATCGCGATTGCAG-3′) and HRML-F primer (5′-GGTGGTTCCTCTAGATCTTCC-3′) were used to sequence the VH and VL chains of the antibody. Sequence analysis of positive clones (O.D.405nm>0.3) was performed by Macrogen (Seoul, Korea).

Next-generation sequencing analysis

A total of 15 sets of phagemid DNA including three initial chicken scFv libraries and three libraries obtained after each of four rounds of bio-panning were analyzed using a MiSeq system (Illumina Inc., San Diego, CA, USA). The MiSeq library for DNA sequencing was prepared using Illumina Nextera XT chemistry (Illumina) following the protocol provided by the manufacturer. The genes from the chicken library were amplified using the forward primer (pre-adaptor, 5′-TCGTCGGCAGCGTC-3′; sequencing primer, 5′-AGATGTGTATAAGAGACAG-3′; specific locus primer, 5′-TCAGCCTCGTCTGCAAGG-3′), and reverse primer (pre-adaptor, 5′-GTCTCGTGGGCTCGG-3′; sequencing primer, 5′-AGATGTGTATAAGAGACAG-3′; specific locus primer, 5′-AGTGGAGGAGACGATGACTTC-3′), respectively. The final libraries were normalized by quantification with LightCycler 480 II (Roche Applied Science, Indianapolis, IN, USA) and qualification with Bioanalyzer (Agilent, Palo Alto, CA, USA). The final loading concentration was adjusted to 11 pM following the MiSeq loading protocol. The MiSeq reagent kit v3 (Illumina) was used for long paired-end reads (2 × 300 bp) sequencing reactions. The sequencing data was processed by CLC Genomics Workbench version 5 (CLC Bio, Aarhus, Denmark) software. Low-quality sequencing data were first trimmed depending on quality scores using PHRED with the minimum quality score of 20 and reads with less than 150 bases in length were discarded[166]. The cleaned-up sequencing data were processed by merging the paired-end sequence reads using fast length adjustment of short reads to obtain complete sequences of the chicken scFv libraries[167]. Sequencing data were further cleaned up using PRINSEQ (San Diego State University, San Diego, CA, USA), setting the minimum quality score at 20 and read length at 150[166]. EMBOSS Needle 6.5.0.0 (The European Bioinformatics Institute (EMBL-EBI), UK) was used to map sequence read in the HCDR3 region, with a threshold score of 300[168]. Subsequently, a custom Perl script was used to summarize and count sequence reads in 15 sets of phagemid DNA. We merged the read counts across all the panning rounds, but for computational and statistical analysis, we only counted the reads existing in the phagemid DNA after the fourth bio-panning round.

Clustering analysis

An optimized number of clusters in the merged sequence read counts was estimated using the clValid algorithm, to facilitate pattern analysis of NGS data for population shifts in antibody clones throughout the bio-panning process[169]. The clValid algorithm validated number of clusters by assessing intra-cluster homogeneity and inter-cluster separation, and the assessment for each and every clustering is represented in the Dunn index[169]. A higher Dunn index indicates better clustering. The 'Internal' cluster validation metrics were chosen, which consider only the data set and the clustering partition, and the intrinsic properties of the data were used to evaluate the quality of the clustering results in designated clustering algorithms such as hierarchical clustering and k-mean clustering[170]. Unsupervised hierarchical clustering analysis was used to cluster HCDR3 sequences according to the number of clusters estimated by clValid. Ward's method was used to measure distances between sequence reads based on read counts throughout the bio-panning, and a heat map visualizing the sequence read changes in each cluster was generated using Gene Pattern v3.9.2 software[171]. Line charts representing the pattern of sequence read changes in each cluster across all the bio-panning rounds were then generated as in a previous study.

Cloning to retrieve scFvs

To rebuild real scFv clones from the virtual HCDR3 sequences in the clusters, we performed two-step linker PCR. In the first PCR step, primers targeting both LFR1-HCDR3 (LFR1_F primer, 5′-GTGGCCCAGGCGGCCCTG-3′) and HCDR3-HFR4 fragments (HFR4_R primer, 5′-CTGGCCGGCCTGGCCACT-3′) were synthesized, based on HCDR3 sequences determined in NGS analysis and phagemid DNA obtained after the 4th round of bio-panning. The second PCR step linked these two gene fragments into a single scFv gene using primers annealing to LFR1 and HFR4 (LFR1_F primer, 5′-GTGGCCCAGGCGGCCCTG-3′; HFR4_R primer, 5′-CTGGCCGGCCTGGCCACT-3′). The scFv gene was ligated into the pComb3XSS phagemid vector and rescued as scFv-displaying phages, as described previously[165]. To measure the binding reactivity of these scFv-displaying phages, we rescued more than 15 clones per HCDR3 sequence, and performed phage enzyme immunoassay as described earlier. We regarded the clone providing the highest optical density at 405 nm as the retrieved clone.

Immunization, Construction of Phage-Displayed scFv Library, and Bio-Panning

White leghorn chickens were immunized and boosted three times with 10 μg of recombinant mouse c-Met-Fc chimera (527-ME; R&D systems, Carlsbad, California, United States). The experiment was approved by the Ethics Committee of BioPOA, Ltd. (ethical approval code: BP-2019-C03-1). One week after the final boosting, total RNA was isolated from spleen, bone marrow, and bursa of Fabricius using TRIzol Reagent (15596018; Invitrogen), and cDNA was synthesized using SuperScript III first-strand cDNA synthesis kit with oligo dT priming (18418020; Invitrogen, Carlsbad, California, United States). Using this cDNA, a phage-displayed scFv library was prepared as described previously[172,173]. V_H and V_L genes were amplified from the cDNA using specific primer sets utilized for the construction of scFv genes. Then, scFv genes were ligated into the pComb3XSS phage display vector, which was transfected into *E. coli* K12 ER2738 cells. Phage-displayed scFv libraries were rescued from transfected cells after infection with VCSM13 helper phage and overnight culture, and then subjected to four rounds of bio-panning using recombinant mouse c-Met (50622-M08H, Sino Biological, Beijing, China)-conjugated magnetic beads (Dynabeads 14302D; Invitrogen). Antigen-coated magnetic beads were washed with 0.05% tween in phosphate-buffered saline (PBS) once for the first round, three times for the second and third rounds, and five times for the fourth round. After each round of bio-panning, phagemid DNA was prepared from bacterial cell pellets using a Qiaprep Spin Miniprep Kit (27104, Qiagen, Hilden, Germany).

Next-Generation Sequencing (NGS)

From five sets of phagemid DNA, short V_H and V_L gene fragments encoding the 3′ part of FR3 and CDR3, and the 5′ part of FR4, were amplified using primers designed to hybridize to FR3 and FR4 of the chicken V_H gene (LFR3: 5′-CCCTTCACGATTCTCCGGTGCC-3′; LFR4: 5′-CTGACCTAGGACGGT CAGGG-3′; HFR3: 5′-GGCTGCAGCTGAACAACCTCAGGGCTG-3′; HFR4: 5′-GGAGGAGACGA TGACTTCGGTCCCGTGG-3′). Other gene fragments encoding the whole V_H and V_L genes were also amplified using specific primers previously described[64]. Prior to NGS analysis, all amplicon libraries were submitted for a quality control procedure on TapeStation 2200 (Agilent Technologies, Santa Clara, California, United States). Libraries having a single peak of correct fragment length were subjected to NGS analysis using the HiSeq 2500 and MiSeq platforms (Illumina, Inc.) for short and whole V_H and V_L gene fragments, respectively. We uploaded the sequence data to NCBI (SRA accession number: PRJNA607865).

To ensure the quality of NGS data, the following pre-processing steps were performed. First, all pair-end reads were merged with PEAR using the developer's default parameters[108]. Second, we filtered out any reads that were compatible with the following description: (1) reads not meeting our minimum quality Phred score, (2) reads not having the primer sequence used in the phage-displayed scFv library construction process, (3) out-of-frame reads, and (4) reads without any identifiable CDR3. The reads were then collated based on their CDR3 sequences and any CDR3 clonotype with read count of less than 2 was discarded.

High-Throughput Clone Retrieval and Phage ELISA

The phagemid library from the final bio-panning round was transfected into *E. coli* K12 ER2738 cells, and then subjected to our high-throughput clonal retrieval procedure using TrueRepertoire (TR) technology, as described previously[63].

The retrieved phage clones were subjected to phage ELISA, as described previously with adequate modifications[173]. Phage clones were rescued overnight from the plate and culture supernatants containing phage that were diluted with equal volumes of 6% bovine serum albumin (BSA) solution in PBS. Phage solutions were then added to microtiter wells (3690, Corning life sciences, Corning, New York, United States) coated with recombinant mouse c-Met or mouse anti-HA antibody (H3663, Merck, Darmstadt, Germany) and blocked with BSA. Microtiter plates were incubated for 2 h at 37 °C and washed three times with 0.05% Tween in PBS, which is followed by 3% BSA in PBS containing horseradish peroxidase (HRP)-conjugated anti-M13 antibody (11973-MM05, Sino Biological) in addition to each well. After incubation and washing as described above, HRP substrate solution 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) (002024, Thermo Fisher Scientific, Waltham, Massachusetts, United States) was added to each well. The plate was incubated 15 min and the absorbance values of each well were measured by a SkanIt microplate reader (Thermo Fisher Scientific) with a fast measurement protocol at a wavelength of 405 nm.

For each clone, the ratio (Relative Absorbance A) of the average absorbance of a recombinant mouse c-Met-coated well vs. an anti-HA antibody-coated well was calculated. The absorbance of an anti-HA-coated well was used to accommodate variations in the amount of phage in each phage clone. We also determined the ratio (Relative Absorbance B) of the average absorbance of a BSA-blocked well to an anti-HA antibody-coated well. When Relative Absorbance A exceeded +3 standard deviation of Relative Absorbance B, we designated the phage clone as antigen-reactive.

Establishment of the Random forest (RF) Models

Random forest (RF), regularized discriminant analysis (RDA), linear discriminant analysis (LDA), support vector machine (SVM), naïve bayes (NB), and AdaBoost (ADA) classification trees were selected for comparison. Our input data for the training of binder prediction models were created using a TR data set consisting of antigen reactivity for scFv clones found in TR analysis and the clonal abundance of their HCDR3 and LCDR3 clonotypes in five sets of phagemid DNA. The caret package for R was used to benchmark popular classification algorithms by their accuracy and Cohen's kappa value. Each algorithm was evaluated across five repetitions of 10-fold cross-validations (50 models in total). This meant that, for each repetition, the training data set was randomly divided into 10 parts and each of the 10 models were cross-validated by one unique part after being trained on the other nine parts of the training data set. No manual tuning was performed during this benchmarking phase[174].

To generate binder prediction models for HCDR3 and LCDR3 clonotypes using a random Forest package, we sampled a proportion of the TR data set without replacement to be used as a training data set for the RF model. The remaining portion of the TR data set served as a validation set to measure the performance of RF models[175]. The following parameters were adjusted to best tune our model's performance: (1) sampling ratio of training data, (2) number of variables (mtry) to randomly sample at each node of the decision-making tree, and (3) number of trees (ntree) to compromise our RF model. We then iterated through all combinations of parameters. Each combination was used to generate 10 different RF models to minimize any biases arising from the training data set not being representative of the TR data set. The validation set was then used to measure the performance of each RF model to determine optimal parameters for the RF. Using the randomForestExplainer package, the minimum depth of each variable was calculated, which is frequently used as a measure of variable importance to elucidate the decision-making process of the algorithm[176]. The minimum depth of a variable is defined as the distance between the root node and the variable's first appearance at a node of the decision tree. Thus, the variable with the smallest mean minimum depth could be regarded as the most important variable. To compare the variable importance results of the prediction model with actual experimental data, we tracked the enrichment pattern by measuring the bio-panning titer and clonal diversity change as Shannon's entropy (SE)[177] by following each round of bio-panning, as described previously[173].

Construction of Antigen-Reactive (AR) and Non-Reactive (NR) Phage-Displayed scFv Library and Phage ELISA

Forty antigen-reactive (AR) and 10 non-reactive (NR) VH and VL genes were chemically synthesized (Twist Bioscience, San Francisco, California, United States). Forty AR VH and 40 AR VL genes were subjected to linker PCR to generate scFv genes, which were used to create the AR phage-displayed scFv library, as described previously20. In a parallel experiment, the NR phage-displayed scFv library was constructed using 10 NR VH and 10 NR VL genes. After a round of bio-panning using recombinant mouse c-Met (50622-M08H; Sino Biological)-conjugated magnetic beads (Dynabeads 14302D; Invitrogen) and washing once with 0.05% tween in PBS, 96 phage clones were randomly rescued from each AR and NR library and subjected to phage ELISA, as described above. After phage ELISA, the nucleotide sequences of scFv clones were determined by Sanger nucleotide sequencing (Macrogen, Seoul, South Korea).

5. Future perspectives

Sequence-based prediction of antigen-antibody interaction

Immense advances in the field of high-throughput antibody repertoire sequencing and screening platform enables the emerging of number of novel antibody discovery technology[178,179]. However, ultimate question: predicting and understanding the vocabulary of antigen-antibody interaction by its' amino acid sequence, remain challenging problems. To the recent, the most accurate method for identifying paratope-epitope interactions is solving the 3D structure of antigen-antibody complexes and determining contact amino acids residues from each binding partners[180]. Cryo-electron microscopy (cryo-EM), X-ray crystallography and NMR (nuclear magnetic resonance) are mostly used methods in analyzing structures of the proteins[181]. From the 3D structure data of antigen-antibody complex, several studies have shown that antigen contacting residues are mostly existed in CDR region, but non-CDR residues are also frequently observed in paratope regions[182,183]. Also, amino acid sequence of the epitopes is indistinguishable from other surface exposed non-epitope residue when the counter-part antibody is not bound as a complex[182,183]. To overcome the redundancy and the low-throughput manner of current approaches in analyzing sequence-based antigen-antibody interactome, machine learning guided computational methods are rapidly advancing.

Current machine learning based approaches have been successful in predicting paratopes[184], epitopes[185], paratope-epitope interaction[126,127,186] and antibody structure[187]. Recent reports have showed slight evidence for the possibility of prediction of antibody-antigen interaction. The antibody repertoire has now established that antibody sequence diversity underlies predictable patterns[188,189]. Also, the presence of universal specificity motifs from different antibodies was recently identified and suggested by showing that high-affinity functional antibodies can be designed by grafting unrelated paratopes

[190]. Further, Akbar et al., showed that structurally identified contact residue has a correlation with somatic hypermutation[191]. This suggests that which somatic hypermutation preserves binding motifs and how a germline would have evolved their interaction motifs. However, without more abundant experimental 3D antibody-antigen interactions data, it is impossible to predict the interactions between antibody and antigen which cannot be crystallized or for unstructured loops of antigens which is generally not existed in structural database. To overcome the certain limitations, the area of *de novo* protein design is rapidly growing owing to advances in artificial intelligent technologies. In prediction of protein structure, *de novo* folding without structural references or comparative modeling from similar template are widely used methods[192]. The recent approaches rely on structural modeling of specific motif of the antigen and antibodies, rather than pre-existing paratope-epitope database[193]. However, high-accuracy prediction of antibody structure is restricted to framework scaffolds, and requires advanced methods in designing hypervariable loop structure in CDR regions[194]. In summary, motif-based prediction of epitope-paratope interactions and structural modeling of antibody frameworks are complementarily developing in prediction of antigen specific antibody sequences.

Machine learning-guided engineering of therapeutic antibodies

Therapeutic antibody candidates have to undergo the optimization process before entering pre-clinical and clinical studies. Assessing immunogenicity, affinity optimization and improving physicochemical properties are the three main goals in antibody optimization works.

Many of clinically available antibodies are derived from immune B cell repertoires of mice, or humanized mice[19,20]. However, non-human antibodies can elicit an immune response, which is known as immunogenicity, and high immunogenicity can influence the efficacy and safety of the mAb therapeutics[19]. To minimize or remove the immunogenicity of mAbs, engineering non-human sequences by substituting the sequence with human Ig germline gene sequence used. These approaches are referred to as de-immunization[195] and humanization[196]. Assessing immunogenicity of the mAbs is time consuming and high-cost process by possessing repetitive and arbitrary generation of mutation library followed by screening of antibody functionality[195]. To reduce the erroneous humanization, accurate validation of immunogenicity or humanness score is needed. Recently, using random forest (RF) algorithm and publicly available antibody sequence database, Marks et al., developed machine learning based validation method measuring humanness scores[197]. However, the non-human sequences in the database were mostly restricted to murine antibody sequence and limitations in the size of the training dataset.

After successful assessment of the immunogenicity of the mAb, optimization of the lead antibodies remains multiple challenges, including productivity (production yield), solubility, thermostability, viscosity, pharmacokinetics[31,198]. Using *in vitro* surface display techniques guided directed evolution, multiple variants could be generated and screened out[34], physicochemical property validation of the variant antibodies should be conducted against the compounds derived from mammalian protein expression system[37]. Low-throughput process including cloning, transfection and repurification limits the scale of the variant screening up to 10^3 clones[37]. Machine learning based prediction of genotype-phenotype relationship has been applied to the engineering of multiple types of proteins[199]. Also, the concepts of machine learning guided directed evolution and generation of mutagenesis library has long been proposed[200,201]. Romero et al. engineered cytochrome enzymes using Gaussian regression models to generate the thermostable variants[202] and Bedbrook et al. designed the variants library of channelrhodopsin protein also utilizing Gaussian regression model[203]. Mason et al. applied the deep learning-based approaches in mammalian cell expressed antibody optimization[179]. Trastuzumab-CDR3 randomized libraries were designed with multiple strategies in generating diverse training data sets, to prevent the overfitting of the model. Similar approaches are expected to be developed in the near future for efficient identification of the most druggable lead compounds.

Single domain antibodies: next-generation therapeutic antibody platform

Back in 1989, stable mouse derived variable heavy chain are isolated and showed binding reactivities to the specific antigens[204]. After that observation, single domain antibodies (sdAbs) were suggested and studied as a unique class of mAbs[204]. Moreover, in 1993, ~ 15 Kda heavy chain only antibodies (HCAbs) was isolated from the sera of camels, which is referred to as VHH[205]. In 1995, ~ 12 Kda sized HCAbs were also isolated from sharks which belongs to cartilaginous fish, and defined as IgNAR (new antigen receptor)[206]. Major advantage of sdAbs is the small size. Compare to full-length Igs, small sized sdAbs can penetrate into hidden epitopes in tissues or infectious pathogens[207]. Also, monomeric structure makes sdAbs effective building blocks for generating multi -valent and -specific antibodies for improvement of therapeutic potency[208].

In next-generation antibody discovery utilizing high-throughput antibody repertoire sequencing and analysis, sdAb repertoires has multiple advantages. To analyze the full-length Ig, heavy chain and light chain genes have to be amplified then subjected into NGS analysis[48]. During that process, low-throughput and high-cost single-cell sequencing platform should be employed to retain the information of natively paired VH-VL origin[209]. In contrast, sdAb repertoire can be annotated with antigen specificity without pairing information, enables the larger depth of the high-throughput analysis. Recently, the increasing success with rapid discovery of sdAbs makes advances in designing the qualified sdAb library. Fridy et al. performed the high-throughput antibody repertoire sequencing combined with mass spectrometric (MS) identification of high-affinity sdAb, from the immunized llamas[210]. Shin et al. designed the improved sdAb library having smaller diversity but better characteristics using autoregressive generative models[211]. Autoregressive generative model was used in synthetic library design by learning alignment-free feature of amino acid sequences having potential binding capacity to diverse target antigens. By utilizing qualified sdAb library and new machine learning algorithms, it is expected to generate a large panel of antibody sequences targeting inaccessible region by the previous approaches.

6. References

1 Cyster, J. G. & Allen, C. D. C. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* **177**, 524-540, doi:10.1016/j.cell.2019.03.016 (2019).

2 Davies DR, M. H. Structural basis of antibody function. *Ann Rev Immunol* **1**, 31 (1983).

3 Talmage, D. W. Immunological Specificity: Unique combinations of selected natural globulins provide an alternative to the classical concept. *Science* **129**, 6 (1959).

4 Xu JL, D. M. Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. *Immunity* **13**, 9 (2000).

5 Mahon, C. M. *et al.* Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. *J Mol Biol* **425**, 1712-1730, doi:10.1016/j.jmb.2013.02.015 (2013).

6 Chen, W., Zhu, Z., Feng, Y. & Dimitrov, D. S. A large human domain antibody library combining heavy and light chain CDR3 diversity. *Mol Immunol* **47**, 912-921, doi:10.1016/j.molimm.2009.09.039 (2010).

7 Murphy, K. W., Casey. Janeway's Immunobiology. *Garland Science, Taylor & Francis Group, LLC* (2016).

8 Schatz, D. G. V(D)J recombination. *Immunol Rev* **200**, 7 (2004).

9 Jung, D., Giallourakis, C., Mostoslavsky, R. & Alt, F. W. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* **24**, 541-570, doi:10.1146/annurev.immunol.23.021704.115830 (2006).

10 Peled, J. U. *et al.* The biochemistry of somatic hypermutation. *Annu Rev Immunol* **26**, 481-511, doi:10.1146/annurev.immunol.26.021607.090236 (2008).

11 Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* **76**, 1-22, doi:10.1146/annurev.biochem.76.061705.090740 (2007).

12 HWANG, J. K., ALT, F. W. & YEAP, L.-S. Related Mechanisms of Antibody Somatic Hypermutation and Class Switch Recombination. *Microbiol Spectr* **3**, doi:10.1128/microbiolspec.MDNA3 (2015).

13 Alberts, B. e. a. Lymphocytes and the Cellular Basis of Adaptive Immunity. *Molecular Biology of the Cell.* **4th edition** (2002).

14 Peterson, M. L. Mechanisms controlling production of membrane and secreted immunoglobulin during B cell development. *Immunol Res* **37**, 14 (2007).

15 Bordon, Y. The many sides of Paul Ehrlich. *Nature Immunology* **17**, S6-S6, doi:10.1038/ni.3601 (2016).

16 Köhler, G., Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 3 (1975).

17 Pirofski, L., Casadevall, A., Rodriguez, L., Zuckier, L. S. & Scharff, M. D. Current state of the hybridoma technology. *Journal of clinical*

immunology **10**, 10 (1990).

18 Liu, J. K. The history of monoclonal antibody development - Progress, remaining challenges and future innovations. *Ann Med Surg (Lond)* **3**, 113-116, doi:10.1016/j.amsu.2014.09.001 (2014).

19 Lu, R. M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *J Biomed Sci* **27**, 1, doi:10.1186/s12929-019-0592-z (2020).

20 Kaplon, H. & Reichert, J. M. Antibodies to watch in 2021. *MAbs* **13**, 1860476, doi:10.1080/19420862.2020.1860476 (2021).

21 Khongorzul, P., Ling, C. J., Khan, F. U., Ihsan, A. U. & Zhang, J. Antibody-Drug Conjugates: A Comprehensive Review. *Mol Cancer Res* **18**, 3-19, doi:10.1158/1541-7786.MCR-19-0582 (2020).

22 Geering, B. & Fussenegger, M. Synthetic immunology: modulating the human immune system. *Trends Biotechnol* **33**, 65-79, doi:10.1016/j.tibtech.2014.10.006 (2015).

23 June, C. H. & Sadelain, M. Chimeric Antigen Receptor Therapy. *N Engl J Med* **379**, 64-73, doi:10.1056/NEJMra1706169 (2018).

24 Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., International Natural Product Sciences, T. & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat Rev Drug Discov* **20**, 200-216, doi:10.1038/s41573-020-00114-z (2021).

25 Wagner, D. L. *et al.* Immunogenicity of CAR T cells in cancer therapy. *Nat Rev Clin Oncol*, doi:10.1038/s41571-021-00476-2 (2021).

26 Feins, S., Kong, W., Williams, E. F., Milone, M. C. & Fraietta, J. A. An introduction to chimeric antigen receptor (CAR) T-cell immunotherapy for human cancer. *Am J Hematol* **94**, S3-S9, doi:10.1002/ajh.25418 (2019).

27 Rafiq, S., Hackett, C. S. & Brentjens, R. J. Engineering strategies to overcome the current roadblocks in CAR T cell therapy. *Nat Rev Clin Oncol* **17**, 147-167, doi:10.1038/s41571-019-0297-y (2020).

28 Huang, R. *et al.* Recent advances in CAR-T cell engineering. *J Hematol Oncol* **13**, 86, doi:10.1186/s13045-020-00910-5 (2020).

29 Newick, K., O'Brien, S., Moon, E. & Albelda, S. M. CAR T Cell Therapy for Solid Tumors. *Annu Rev Med* **68**, 139-152, doi:10.1146/annurev-med-062315-120245 (2017).

30 Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203-214, doi:10.1038/nrd3078 (2010).

31 Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci U S A* **114**, 944-949, doi:10.1073/pnas.1616408114 (2017).

32 Sun, A. & Benet, L. Z. Late-Stage Failures of Monoclonal Antibody Drugs: A Retrospective Case Study Analysis. *Pharmacology* **105**, 145-163, doi:10.1159/000505379 (2020).

33 Zaroff S, T. G. Hybridoma technology: the preferred method for monoclonal antibody generation for in vivo applications. *BioTechniques* **67**, 3 (2019).

34 Bradbury, A. R., Sidhu, S., Dubel, S. & McCafferty, J. Beyond natural antibodies: the power of in vitro display technologies. *Nat Biotechnol* **29**, 245-254, doi:10.1038/nbt.1791 (2011).

35 Shim, H. Synthetic approach to the generation of antibody diversity. *BMB Rep* **48**, 489-494, doi:10.5483/bmbrep.2015.48.9.120 (2015).

36 GP, S. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 3 (1985).

37 Horlick, R. A. *et al.* Simultaneous surface display and secretion of proteins from mammalian cells facilitate efficient in vitro selection and maturation of antibodies. *J Biol Chem* **288**, 19861-19869, doi:10.1074/jbc.M113.452482 (2013).

38 Nixon, A. E., Sexton, D. J. & Ladner, R. C. Drugs derived from phage display: from candidate identification to clinical practice. *MAbs* **6**, 73-85, doi:10.4161/mabs.27240 (2014).

39 Boder ET, W. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* **15**, 5 (1997).

40 Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* **1**, 755-768, doi:10.1038/nprot.2006.94 (2006).

41 Yamashita, M., Katakura, Y. & Shirahata, S. Recent advances in the generation of human monoclonal antibody. *Cytotechnology* **55**, 55-60, doi:10.1007/s10616-007-9072-5 (2007).

42 LL, G. Transgenic Mouse Strains as Platforms for the Successful Discovery and Development of Human Therapeutic Monoclonal Antibodies. *Current Drug Discovery Technologies* **11**, 74 (2014).

43 Lee, E. C. *et al.* Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nat Biotechnol* **32**, 356-363, doi:10.1038/nbt.2825 (2014).

44 Murphy, A. J. *et al.* Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. *Proc Natl Acad Sci U S A* **111**, 5153-5158, doi:10.1073/pnas.1324022111 (2014).

45 Macdonald, L. E. *et al.* Precise and in situ genetic humanization of 6 Mb of mouse immunoglobulin genes. *Proc Natl Acad Sci U S A* **111**, 5147-5152, doi:10.1073/pnas.1323896111 (2014).

46 Chen, W. C. & Murawsky, C. M. Strategies for Generating Diverse Antibody Repertoires Using Transgenic Animals Expressing Human Antibodies. *Front Immunol* **9**, 460, doi:10.3389/fimmu.2018.00460 (2018).

47 Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* **17**, 53, doi:10.1186/s13059-016-0917-0 (2016).

48 Rouet, R., Jackson, K. J. L., Langley, D. B. & Christ, D. Next-Generation Sequencing of Antibody Display Repertoires. *Front Immunol* **9**, 118, doi:10.3389/fimmu.2018.00118 (2018).

49 Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* **32**, 158-168, doi:10.1038/nbt.2782 (2014).

50 Kovaltsuk, A. *et al.* Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLoS Comput Biol* **16**, e1007636, doi:10.1371/journal.pcbi.1007636 (2020).

51 Jardine JG, K. D., Havenar-Daughton C, Sarkar A, Briney B, Sok D, et al. HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science* **351**, 6.

52 Kim, S. I., Noh, J., Kim, S., Choi, Y., Yoo, D.K., Lee, Y., Lee, H., Jung, J., Kang, C.K., Song, K.H., et al. Stereotypic neutralizing VH antibodies against SARS-CoV-2 spike protein receptor binding domain in patients with COVID-19 and healthy individuals. *Sci Transl Med* **13**, eabd6990 (2021).

53 DeFalco, J. *et al.* Non-progressing cancer patients have persistent B cell responses expressing shared antibody paratopes that target public tumor antigens. *Clin Immunol* **187**, 37-45, doi:10.1016/j.clim.2017.10.002 (2018).

54 Benjamin, J. E. *et al.* First-in-human phase Ib study of ATRC-101, an engineered version of a patient-derived antibody targeting a tumor-restricted ribonucleoprotein complex. *Journal of Clinical Oncology* **38**, TPS3168-TPS3168, doi:10.1200/JCO.2020.38.15_suppl.TPS3168 (2020).

55 Kim, S. I. *et al.* Generation of a Nebulizable CDR-Modified MERS-CoV Neutralizing Human Antibody. *Int J Mol Sci* **20**, doi:10.3390/ijms20205073 (2019).

56 Lloyd, C. *et al.* Modelling the human immune response: performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. *Protein Eng Des Sel* **22**, 159-168, doi:10.1093/protein/gzn058 (2009).

57 Schwimmer, L. J. *et al.* Discovery of diverse and functional antibodies from large human repertoire antibody libraries. *J Immunol Methods* **391**, 60-71, doi:10.1016/j.jim.2013.02.010 (2013).

58 Weber, M. *et al.* A highly functional synthetic phage display library containing over 40 billion human antibody clones. *PLoS One* **9**, e100000, doi:10.1371/journal.pone.0100000 (2014).

59 Asensio, M. A. *et al.* Antibody repertoire analysis of mouse immunization protocols using microfluidics and molecular genomics. *MAbs* **11**, 870-883, doi:10.1080/19420862.2019.1583995 (2019).

60 Feng, R. *et al.* Isolation of rabbit single domain antibodies to B7-H3 via protein immunization and phage display. *Antib Ther* **3**, 10-17, doi:10.1093/abt/tbaa002 (2020).

61 Wege, A. K. *et al.* A novel rabbit derived anti-HER2 antibody with pronounced therapeutic effectiveness on HER2-positive breast cancer cells in vitro and in humanized tumor mice (HTM). *J Transl Med* **18**, 316, doi:10.1186/s12967-020-02484-9 (2020).

62 Kim, Y. J., Lebreton, F., Kaiser, C., Cruciere, C. & Remond, M. Isolation of foot-and-mouth disease virus specific bovine antibody fragments from phage display libraries. *J Immunol Methods* **286**, 155-166, doi:10.1016/j.jim.2004.01.002 (2004).

63 Noh, J. *et al.* High-throughput retrieval of physical DNA for NGS-identifiable clones in phage display library. *MAbs* **11**, 532-545, doi:10.1080/19420862.2019.1571878 (2019).

64 Yang, W. *et al.* Next-generation sequencing enables the discovery of more diverse positive clones from a phage-displayed antibody library. *Exp Mol Med* **49**, e308, doi:10.1038/emm.2017.22 (2017).

65 Feng, M. *et al.* Construction and next-generation sequencing analysis of a large phage-displayed VNAR single-domain antibody library from six naive nurse sharks. *Antib Ther* **2**, 1-11, doi:10.1093/abt/tby011 (2019).

66 Lefranc, M. P. & Lefranc, G. Immunoglobulins or Antibodies: IMGT((R)) Bridging Genes, Structures and Functions. *Biomedicines* **8**, doi:10.3390/biomedicines8090319 (2020).

67 Okino, S. T., Kong, M., Sarras, H. & Wang, Y. Evaluation of bias associated with high-multiplex, target-specific pre-amplification. *Biomol Detect Quantif* **6**, 13-21, doi:10.1016/j.bdq.2015.12.001 (2016).

68 Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol* **17**, 61, doi:10.1186/s12896-017-0379-9 (2017).

69 Markoulatos, P., Siafakas, N. & Moncany, M. Multiplex polymerase chain reaction: a practical approach. *J Clin Lab Anal* **16**, 47-51, doi:10.1002/jcla.2058 (2002).

70 Rowlands, V. *et al.* Optimisation of robust singleplex and multiplex droplet digital PCR assays for high confidence mutation detection in circulating tumour DNA. *Sci Rep* **9**, 12620, doi:10.1038/s41598-019-49043-x (2019).

71 van Dongen, J. J. *et al.* Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257-2317, doi:10.1038/sj.leu.2403202 (2003).

72 Charlton, K. A., Moyle, S., Porter, A. J. & Harris, W. J. Analysis of the diversity of a sheep antibody repertoire as revealed from a bacteriophage display library. *J Immunol* **164**, 6221-6229, doi:10.4049/jimmunol.164.12.6221 (2000).

73 Dong, S., Bo, Z., Zhang, C., Feng, J. & Liu, X. Screening for single-chain variable fragment antibodies against multiple Cry1 toxins from an immunized mouse phage display antibody library. *Appl Microbiol Biotechnol* **102**, 3363-3374, doi:10.1007/s00253-018-8797-8 (2018).

74 Dooley, H. Selection and characterization of naturally occurring single-domain (IgNAR) antibody fragments from immunized sharks by phage display. *Molecular Immunology* **40**, 25-33, doi:10.1016/s0161-5890(03)00084-1 (2003).

75 Harmsen, M. M., Van Solt, C. B., Fijten, H. P. & Van Setten, M. C. Prolonged in vivo residence times of llama single-domain antibody fragments in pigs by binding to porcine immunoglobulins. *Vaccine* **23**,

4926-4934, doi:10.1016/j.vaccine.2005.05.017 (2005).

76 Lim, T. S. & Chan, S. K. Immune Antibody Libraries: Manipulating The Diverse Immune Repertoire for Antibody Discovery. *Curr Pharm Des* **22**, 10, doi:10.2174/1381612822666160920 (2016).

77 Pelat, T., Hust, M. & Thullier, P. Obtention and engineering of non-human primate (NHP) antibodies for therapeutics. *Mini Rev Med Chem* **9**, 6 (2009).

78 Saggy, I. *et al.* Antibody isolation from immunized animals: comparison of phage display and antibody discovery via V gene repertoire mining. *Protein Eng Des Sel* **25**, 539-549, doi:10.1093/protein/gzs060 (2012).

79 Xu, C. *et al.* Construction of an Immunized Rabbit Phage Display Library for Selecting High Activity against Bacillus thuringiensis Cry1F Toxin Single-Chain Antibodies. *J Agric Food Chem* **65**, 6016-6022, doi:10.1021/acs.jafc.7b01985 (2017).

80 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).

81 Peng, H. *et al.* Mining Naive Rabbit Antibody Repertoires by Phage Display for Monoclonal Antibodies of Therapeutic Utility. *J Mol Biol* **429**, 2954-2973, doi:10.1016/j.jmb.2017.08.003 (2017).

82 Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393-397, doi:10.1038/s41586-019-0879-y (2019).

83 Kim, S., I. *et al.* Stereotypic neutralizing VH antibodies against SARS-CoV-2 spike protein receptor binding domain in COVID-19 patients and healthy individuals. *Sci Transl Med* **eabd6990**, doi:10.1126/scitranslmed.abd6990 (2021).

84 Turchaninova, M. A. *et al.* High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* **11**, 1599-1616, doi:10.1038/nprot.2016.093 (2016).

85 Barennes, P. *et al.* Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat Biotechnol*, doi:10.1038/s41587-020-0656-3 (2020).

86 Galson, J. D. *et al.* In-Depth Assessment of Within-Individual and Inter-Individual Variation in the B Cell Receptor Repertoire. *Front Immunol* **6**, 531, doi:10.3389/fimmu.2015.00531 (2015).

87 Riedel, R. *et al.* Discrete populations of isotype-switched memory B lymphocytes are maintained in murine spleen and bone marrow. *Nat Commun* **11**, 2570, doi:10.1038/s41467-020-16464-6 (2020).

88 Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398-402, doi:10.1038/s41586-019-0934-8 (2019).

89 Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* **7**, 121, doi:10.1186/s13073-015-0243-2 (2015).

90 Chao, A. *et al.* Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol Monogr* **84**, 23 (2014).

91 Kim, S. *et al.* Efficient Selection of Antibodies Reactive to Homologous Epitopes on Human and Mouse Hepatocyte Growth Factors by Next-Generation Sequencing-Based Analysis of the B Cell Repertoire. *Int J Mol Sci* **20**, doi:10.3390/ijms20020417 (2019).

92 Cao, Y. *et al.* Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* **182**, 73-84 e16, doi:10.1016/j.cell.2020.05.025 (2020).

93 Henry Dunand, C. J. & Wilson, P. C. Restricted, canonical, stereotyped and convergent immunoglobulin responses. *Philos Trans R Soc Lond B Biol Sci* **370**, doi:10.1098/rstb.2014.0238 (2015).

94 Seiffert, M. Antibody peptides as cancer vaccine - turning weapons to targets. *Clin Cancer Res*, doi:10.1158/1078-0432.CCR-20-3977 (2020).

95 Snir, O. *et al.* Stereotyped antibody responses target posttranslationally modified gluten in celiac disease. *JCI Insight* **2**, doi:10.1172/jci.insight.93961 (2017).

96 Tonouchi, K. *et al.* Stereotyped B-cell response that counteracts antigenic variation of influenza viruses. *Int Immunol* **32**, 613-621, doi:10.1093/intimm/dxaa038 (2020).

97 Robbiani, D. F. *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437-442, doi:10.1038/s41586-020-2456-9 (2020).

98 Stamatatos, L., Morris, L., Burton, D. R. & Mascola, J. R. Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nat Med* **15**, 866-870, doi:10.1038/nm.1949 (2009).

99 Hosie, M. J., Pajek, D., Samman, A. & Willett, B. J. Feline immunodeficiency virus (FIV) neutralization: a review. *Viruses* **3**, 1870-1890, doi:10.3390/v3101870 (2011).

100 Dustin, L. B., Cashman, S. B. & Laidlaw, S. M. Immune control and failure in HCV infection--tipping the balance. *J Leukoc Biol* **96**, 535-548, doi:10.1189/jlb.4RI0214-126R (2014).

101 Szczepanek, S. M., Barrette, R. W., Rood, D., Alejo, D. & Silbart, L. K. Xenoepitope substitution avoids deceptive imprinting and broadens the immune response to foot-and-mouth disease virus. *Clin Vaccine Immunol* **19**, 461-467, doi:10.1128/CVI.00035-12 (2012).

102 Du, L. *et al.* Introduction of neutralizing immunogenicity index to the rational design of MERS coronavirus subunit vaccines. *Nat Commun* **7**, 13473, doi:10.1038/ncomms13473 (2016).

103 Yu, L. *et al.* Critical epitopes in the nucleocapsid protein of SFTS virus recognized by a panel of SFTS patients derived human monoclonal antibodies. *PLoS One* **7**, e38291, doi:10.1371/journal.pone.0038291 (2012).

104 Thaa, B., Sinhadri, B. C., Tielesch, C., Krause, E. & Veit, M. Signal

peptide cleavage from GP5 of PRRSV: a minor fraction of molecules retains the decoy epitope, a presumed molecular cause for viral persistence. *PLoS One* **8**, e65548, doi:10.1371/journal.pone.0065548 (2013).

105 Gillet, L., May, J. S., Colaco, S. & Stevenson, P. G. The murine gammaherpesvirus-68 gp150 acts as an immunogenic decoy to limit virion neutralization. *PLoS One* **2**, e705, doi:10.1371/journal.pone.0000705 (2007).

106 Trible, B. R. *et al.* Antibody recognition of porcine circovirus type 2 capsid protein epitopes after vaccination, infection, and disease. *Clin Vaccine Immunol* **18**, 749-757, doi:10.1128/CVI.00418-10 (2011).

107 Yoo, D. K. *et al.* Machine Learning-Guided Prediction of Antigen-Reactive In Silico Clonotypes Based on Changes in Clonal Abundance through Bio-Panning. *Biomolecules* **10**, doi:10.3390/biom10030421 (2020).

108 Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620, doi:10.1093/bioinformatics/btt593 (2014).

109 McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**, W20-25, doi:10.1093/nar/gkh435 (2004).

110 Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* **12**, 380-381, doi:10.1038/nmeth.3364 (2015).

111 Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* **41**, W34-40, doi:10.1093/nar/gkt382 (2013).

112 Hsieh, T. C., Ma, K. H., Chao, A. & McInerny, G. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* **7**, 1451-1456, doi:10.1111/2041-210x.12613 (2016).

113 Meng, L. *et al.* Use of Exome Sequencing for Infants in Intensive Care Units: Ascertainment of Severe Single-Gene Disorders and Effect on Medical Management. *JAMA Pediatr* **171**, e173438, doi:10.1001/jamapediatrics.2017.3438 (2017).

114 Comoglio, P. M., Trusolino, L. & Boccaccio, C. Known and novel roles of the MET oncogene in cancer: a coherent approach to targeted therapy. *Nat Rev Cancer* **18**, 341-358, doi:10.1038/s41568-018-0002-y (2018).

115 Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* **15**, 81-94, doi:10.1038/nrclinonc.2017.166 (2018).

116 Park, J. *et al.* FOXO1 Suppression is a Determinant of Acquired Lapatinib-Resistance in HER2-Positive Gastric Cancer Cells Through MET Upregulation. *Cancer Res Treat* **50**, 239-254, doi:10.4143/crt.2016.580 (2018).

117 Pietrantonio, F. *et al.* Biomarkers of Primary Resistance to Trastuzumab in HER2-Positive Metastatic Gastric Cancer Patients:

the AMNESIA Case-Control Study. *Clin Cancer Res* **24**, 1082-1089, doi:10.1158/1078-0432.CCR-17-2781 (2018).

118 Lai, G. G. L., T.H. Lim, J. Liew, P.J. Kwang, X.L. Nahar, R. Aung, Z.W. Takano, A. Lee, Y.Y. Lau, D.P. et al. Clonal MET Amplification as a Determinant of Tyrosine Kinase Inhibitor Resistance in Epidermal Growth Factor Receptor–Mutant Non–Small-Cell Lung Cancer. *J. Clin. Oncol.* **37**, 9, doi:10.1200/JCO.18 (2019).

119 Martin, V. *et al.* Met inhibition revokes IFNgamma-induction of PD-1 ligands in MET-amplified tumours. *Br J Cancer* **120**, 527-536, doi:10.1038/s41416-018-0315-3 (2019).

120 Saigi, M. *et al.* MET-Oncogenic and JAK2-Inactivating Alterations Are Independent Factors That Affect Regulation of PD-L1 Expression in Lung Cancer. *Clin Cancer Res* **24**, 4579-4587, doi:10.1158/1078-0432.CCR-18-0267 (2018).

121 Glodde, N. *et al.* Reactive Neutrophil Responses Dependent on the Receptor Tyrosine Kinase c-MET Limit Cancer Immunotherapy. *Immunity* **47**, 789-802 e789, doi:10.1016/j.immuni.2017.09.012 (2017).

122 Papaccio, F. *et al.* HGF/MET and the Immune System: Relevance for Cancer Immunotherapy. *Int J Mol Sci* **19**, doi:10.3390/ijms19113595 (2018).

123 Kim, S. T. *et al.* First-in-human phase I trial of anti-hepatocyte growth factor antibody (YYB101) in refractory solid tumor patients. *Ther Adv Med Oncol* **12**, 1758835920926796, doi:10.1177/1758835920926796 (2020).

124 Rolfo, C., Van Der Steen, N., Pauwels, P. & Cappuzzo, F. Onartuzumab in lung cancer: the fall of Icarus? *Expert Rev Anticancer Ther* **15**, 487-489, doi:10.1586/14737140.2015.1031219 (2015).

125 Parola, C., Neumeier, D. & Reddy, S. T. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. *Immunology* **153**, 31-41, doi:10.1111/imm.12838 (2018).

126 Brown, A. J. *et al.* Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *Molecular Systems Design & Engineering* **4**, 701-736, doi:10.1039/c9me00071b (2019).

127 Norman, R. A. *et al.* Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief Bioinform* **21**, 1549-1567, doi:10.1093/bib/bbz095 (2020).

128 Sing A, T. N., Sharma A. A review of supervised machine learning algorithms. *IEEE* (2016).

129 Kulkarni, V. Y. S., P.K. Effective Learning and Classification using Random Forest Algorithm. *Int. J. Eng. Innov. Technolgy* **3**, 7 (2014).

130 Hughes, V. S. & Siemann, D. W. Have Clinical Trials Properly Assessed c-Met Inhibitors? *Trends Cancer* **4**, 94-97, doi:10.1016/j.trecan.2017.11.009 (2018).

131 Moosavi, F., Giovannetti, E., Saso, L. & Firuzi, O. HGF/MET pathway aberrations as diagnostic, prognostic, and predictive biomarkers in

human cancers. *Crit Rev Clin Lab Sci* 56, 533-566, doi:10.1080/10408363.2019.1653821 (2019).

132 Kim, K. H. & Kim, H. Progress of antibody-based inhibitors of the HGF-cMET axis in cancer therapy. *Exp Mol Med* 49, e307, doi:10.1038/emm.2017.17 (2017).

133 Oh, Y. M. S., Y.-J.; Lee, S.B.; Jeong, Y.; Kim, B.; Kim, G.W.; Kim, K.E.; Lee, J.M.; Cho, M.-Y.; Choi, J.; et al. A New Anti-c-Met Antibody Selected by a Mechanism-Based Dual-Screening Method: Therapeutic Potential in Cancer. *Mol. Cells* 34, 7 (2012).

134 Patnaik, A. *et al.* A phase I study of LY3164530, a bispecific antibody targeting MET and EGFR, in patients with advanced or metastatic cancer. *Cancer Chemother Pharmacol* 82, 407-418, doi:10.1007/s00280-018-3623-7 (2018).

135 Park, K. *et al.* OA10.06 A First-in-Human Phase 1 Trial of the EGFR-cMET Bispecific Antibody JNJ-61186372 in Patients with Advanced Non-Small Cell Lung Cancer (NSCLC). *Journal of Thoracic Oncology* 13, S344-S345, doi:10.1016/j.jtho.2018.08.292 (2018).

136 Pierpont, T. M., Limper, C. B. & Richards, K. L. Past, Present, and Future of Rituximab-The World's First Oncology Monoclonal Antibody Therapy. *Front Oncol* 8, 163, doi:10.3389/fonc.2018.00163 (2018).

137 Zahavi, D., AlDeghaither, D., O'Connell, A. & Weiner, L. M. Enhancing antibody-dependent cell-mediated cytotoxicity: a strategy for improving antibody-based immunotherapy. *Antibody Therapeutics* 1, 7-12, doi:10.1093/abt/tby002 (2018).

138 Saffi, G. T. & Botelho, R. J. Lysosome Fission: Planning for an Exit. *Trends Cell Biol* 29, 635-646, doi:10.1016/j.tcb.2019.05.003 (2019).

139 Rinnerthaler, G., Gampenrieder, S. P. & Greil, R. HER2 Directed Antibody-Drug-Conjugates beyond T-DM1 in Breast Cancer. *Int J Mol Sci* 20, doi:10.3390/ijms20051115 (2019).

140 Romero, D. Haematological cancer: Blinatumomab facilitates complete responses. *Nat Rev Clin Oncol* 15, 200, doi:10.1038/nrclinonc.2018.24 (2018).

141 Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody-antigen recognition. *Front Immunol* 4, 302, doi:10.3389/fimmu.2013.00302 (2013).

142 Jespersen, M. C., Mahajan, S., Peters, B., Nielsen, M. & Marcatili, P. Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes. *Front Immunol* 10, 298, doi:10.3389/fimmu.2019.00298 (2019).

143 Smith GP, P. V. Phage Display. *Chem. Rev.* 97, 10 (1997).

144 Mimmi, S., Maisano, D., Quinto, I. & Iaccino, E. Phage Display: An Overview in Context to Drug Discovery. *Trends Pharmacol Sci* 40, 87-91, doi:10.1016/j.tips.2018.12.005 (2019).

145 Peltomaa, R., Benito-Pena, E., Barderas, R. & Moreno-Bondi, M. C. Phage Display in the Quest for New Selective Recognition Elements for Biosensors. *ACS Omega* 4, 11569-11580, doi:10.1021/acsomega.9b01206 (2019).

146 Kaplon, H., Muralidharan, M., Schneider, Z. & Reichert, J. M. Antibodies to watch in 2020. *MAbs* **12**, 1703531, doi:10.1080/19420862.2019.1703531 (2020).

147 Ravn, U. *et al.* By-passing in vitro screening--next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* **38**, e193, doi:10.1093/nar/gkq789 (2010).

148 Ravn, U. *et al.* Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60**, 99-110, doi:10.1016/j.ymeth.2013.03.001 (2013).

149 D'Angelo, S. *et al.* From deep sequencing to actual clones. *Protein Eng Des Sel* **27**, 301-307, doi:10.1093/protein/gzu032 (2014).

150 Hu, D. *et al.* Effective Optimization of Antibody Affinity by Phage Display Integrated with High-Throughput DNA Synthesis and Sequencing Technologies. *PLoS One* **10**, e0129125, doi:10.1371/journal.pone.0129125 (2015).

151 Spiliotopoulos, A. *et al.* Sensitive recovery of recombinant antibody clones after their in silico identification within NGS datasets. *J Immunol Methods* **420**, 50-55, doi:10.1016/j.jim.2015.03.005 (2015).

152 Miyazaki, N. K., N.; Akazawa, Y.; Takashima, M.; Hagihara, Y.; Inoue, N.; Matsuda, T.; Ogawa, R.;Inoue, S.; Ito, Y. Isolation and characterization of antigen-specific alpaca (Lama pacos) VHH antibodies by biopanning followed by high-throughput sequencing. *J. Biochem* **158**, 11 (2015).

153 Lovgren, J., Pursiheimo, J. P., Pyykko, M., Salmi, J. & Lamminmaki, U. Next generation sequencing of all variable loops of synthetic single framework scFv-Application in anti-HDL antibody selections. *N Biotechnol* **33**, 790-796, doi:10.1016/j.nbt.2016.07.009 (2016).

154 Barreto, K. *et al.* Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. *Nucleic Acids Res* **47**, e50, doi:10.1093/nar/gkz131 (2019).

155 Lowden, M. J. H., K.A. Oxford nanopore sequencing enables rapid discovery of single-domain antibodies from phage display libraries. *Biotechniques* **65**, 6 (2018).

156 Mei, M. *et al.* Application of modified yeast surface display technologies for non-Antibody protein engineering. *Microbiol Res* **196**, 118-128, doi:10.1016/j.micres.2016.12.002 (2017).

157 Sha, S., Agarabi, C., Brorson, K., Lee, D. Y. & Yoon, S. N-Glycosylation Design and Control of Therapeutic Monoclonal Antibodies. *Trends Biotechnol* **34**, 835-846, doi:10.1016/j.tibtech.2016.02.013 (2016).

158 Sydow, J. F. *et al.* Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. *PLoS One* **9**, e100736, doi:10.1371/journal.pone.0100736 (2014).

159 Tomar, D. S. *et al.* In-silico prediction of concentration-dependent viscosity curves for monoclonal antibody solutions. *MAbs* **9**, 476-489, doi:10.1080/19420862.2017.1285479 (2017).

160 Obrezanova, O. *et al.* Aggregation risk prediction for antibodies and its application to biotherapeutic development. *MAbs* **7**, 352-363, doi:10.1080/19420862.2015.1007828 (2015).

161 Sankar, K. *et al.* Prediction of methionine oxidation risk in monoclonal antibodies using a machine learning method. *MAbs* **10**, 1281-1290, doi:10.1080/19420862.2018.1518887 (2018).

162 Mason, D. M. *et al.* Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *bioRxiv*, doi:10.1101/617860 (2019).

163 Liu, G. *et al.* Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126-2133, doi:10.1093/bioinformatics/btz895 (2020).

164 Bujotzek, A. *et al.* VH-VL orientation prediction for antibody humanization candidate selection: A case study. *MAbs* **8**, 288-305, doi:10.1080/19420862.2015.1117720 (2016).

165 Han, J. *et al.* A phosphorylation pattern-recognizing antibody specifically reacts to RNA polymerase II bound to exons. *Exp Mol Med* **48**, e271, doi:10.1038/emm.2016.101 (2016).

166 Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864, doi:10.1093/bioinformatics/btr026 (2011).

167 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

168 Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* **43**, W580-584, doi:10.1093/nar/gkv279 (2015).

169 Brock, G. P., V; Datta, S; Datta, S. clValid: An R Package for Cluster Validation. *J Stat Softw* **25**, 22 (2008).

170 Hartigan, J. W. M. A k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* **28**, 9 (1979).

171 Kuehn, H., Liberzon, A., Reich, M. & Mesirov, J. P. Using GenePattern for gene expression analysis. *Curr Protoc Bioinformatics* **Chapter 7**, Unit 7 12, doi:10.1002/0471250953.bi0712s22 (2008).

172 Andris-Widhopf, J. R., C. Steinberger, P. Fuller, R. Barbas, C.F. Methods for the generation of chicken monoclonal antibody fragments by phage display. *J. Immunol. Methods* **242**, 23 (2000).

173 Barbas, C. I. B. D., Scott JK, Silverman G.J. Phage Display: A Laboratory Manual. *Cold Spring Harbor Laboratory Press* (2001).

174 Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw* **28** (2008).

175 Liaw A, W. M. Classification and regression by randomforest. *R News* **2**, 5 (2002).

176 Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, doi:10.18637/jss.v077.i01 (2017).

177 Rempala, G. A. & Seweryn, M. Methods for diversity and overlap

analysis in T-cell receptor populations. *J Math Biol* **67**, 1339-1368, doi:10.1007/s00285-012-0589-7 (2013).

178 Graves, J. *et al.* A Review of Deep Learning Methods for Antibodies. *Antibodies (Basel)* **9**, doi:10.3390/antib9020012 (2020).

179 Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5**, 600-612, doi:10.1038/s41551-021-00699-9 (2021).

180 Van Regenmortel, M. H. Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. *J Mol Recognit* **27**, 627-639, doi:10.1002/jmr.2394 (2014).

181 Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* **20**, 681-697, doi:10.1038/s41580-019-0163-x (2019).

182 Greiff, V., Yaari, G. & Cowell, L. G. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology* **24**, 109-119, doi:10.1016/j.coisb.2020.10.010 (2020).

183 Mahajan, S. *et al.* Benchmark datasets of immune receptor-epitope structural complexes. *BMC Bioinformatics* **20**, 490, doi:10.1186/s12859-019-3109-6 (2019).

184 Deac, A., VeliCkovic, P. & Sormanni, P. Attentive Cross-Modal Paratope Prediction. *J Comput Biol* **26**, 536-545, doi:10.1089/cmb.2018.0175 (2019).

185 Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* **8**, e1002829, doi:10.1371/journal.pcbi.1002829 (2012).

186 Raybould, M. I. J., Wong, W. K. & Deane, C. M. Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Molecular Systems Design & Engineering* **4**, 679-688, doi:10.1039/c9me00034h (2019).

187 Weitzner, B. D. *et al.* Modeling and docking of antibody structures with Rosetta. *Nat Protoc* **12**, 401-416, doi:10.1038/nprot.2016.180 (2017).

188 Greiff, V. *et al.* Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *J Immunol* **199**, 2985-2997, doi:10.4049/jimmunol.1700594 (2017).

189 Greiff, V. *et al.* Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep* **19**, 1467-1478, doi:10.1016/j.celrep.2017.04.054 (2017).

190 Nimrod, G. *et al.* Computational Design of Epitope-Specific Functional Antibodies. *Cell Rep* **25**, 2121-2131 e2125, doi:10.1016/j.celrep.2018.10.081 (2018).

191 Akbar, R. *et al.* A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep* **34**, 108856, doi:10.1016/j.celrep.2021.108856 (2021).

192 Bender, B. J. *et al.* Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry* **55**, 4748-4763, doi:10.1021/acs.biochem.6b00444 (2016).

193 Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320-327, doi:10.1038/nature19946 (2016).

194 Norn, C. H., Lapidoth, G. & Fleishman, S. J. High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments. *Proteins* **85**, 30-38, doi:10.1002/prot.25185 (2017).

195 Zinsli, L. V., Stierlin, N., Loessner, M. J. & Schmelcher, M. Deimmunization of protein therapeutics - Recent advances in experimental and computational epitope prediction and deletion. *Comput Struct Biotechnol J* **19**, 315-329, doi:10.1016/j.csbj.2020.12.024 (2021).

196 Safdari, Y., Farajnia, S., Asgharzadeh, M. & Khalili, M. Antibody humanization methods - a review and update. *Biotechnol Genet Eng Rev* **29**, 175-186, doi:10.1080/02648725.2013.801235 (2013).

197 Marks, C., Hummer, A. M., Chin, M. & Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, doi:10.1093/bioinformatics/btab434 (2021).

198 Sharma, V. K. *et al.* In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. *Proc Natl Acad Sci U S A* **111**, 18601-18606, doi:10.1073/pnas.1421779112 (2014).

199 Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A* **116**, 8852-8858, doi:10.1073/pnas.1901979116 (2019).

200 Fox, R. Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J Theor Biol* **234**, 187-199, doi:10.1016/j.jtbi.2004.11.031 (2005).

201 Fox, R. *et al.* Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng* **16**, 589-597, doi:10.1093/protein/gzg077 (2003).

202 Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci U S A* **110**, E193-201, doi:10.1073/pnas.1215251110 (2013).

203 Bedbrook, C. N., Yang, K. K., Rice, A. J., Gradinaru, V. & Arnold, F. H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput Biol* **13**, e1005786, doi:10.1371/journal.pcbi.1005786 (2017).

204 Holliger, P. & Hudson, P. J. Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* **23**, 1126-1136, doi:10.1038/nbt1142 (2005).

205 Hamers-Casterman, C. A., T; Muyldermans, S; Robinson, G; Hamers,

C; Bajyana Songa, E; Bendahman, N; Hamers, R. Naturally occuring antibodies devoid of light chains. *Nature* **363**, 3 (1993).

206 Greenberg, A. A., D; Hughes, M; Hughes, A; Mckinney, EC; Flajnik, MF. A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature* **374**, 6 (1995).

207 Samaranayake, H., Wirth, T., Schenkwein, D., Raty, J. K. & Yla-Herttuala, S. Challenges in monoclonal antibody-based therapies. *Ann Med* **41**, 322-331, doi:10.1080/07853890802698842 (2009).

208 Kaiser, P. D., Maier, J., Traenkle, B., Emele, F. & Rothbauer, U. Recent progress in generating intracellular functional antibody fragments to target and trace cellular components in living cells. *Biochim Biophys Acta* **1844**, 1933-1942, doi:10.1016/j.bbapap.2014.04.019 (2014).

209 Goldstein, L. D. *et al.* Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun Biol* **2**, 304, doi:10.1038/s42003-019-0551-y (2019).

210 Fridy, P. C. *et al.* A robust pipeline for rapid production of versatile nanobody repertoires. *Nat Methods* **11**, 1253-1260, doi:10.1038/nmeth.3170 (2014).

211 Shin, J.-E. *et al.* Protein Design and Variant Prediction Using Autoregressive 1 Generative Models. *bioRxiv*, doi:10.1101/757252 (2021).

초록

연구의 배경: 단일 클론 항체 (monoclonal antibody, mAb) 는 B 세포에서 생산되어 표적 항원에 특이적으로 결합하는 폴리펩타이드 복합체 이다. 분자 및 세포 클로닝 기술의 발전으로 재조합 단일 클론 항체를 대용량으로 생산하는것이 가능해졌으며, 이를 바탕으로 다양한 연구 및 임상 분야에서의 활용이 확대되고 있다. 또한 치료용 항체를 효율적으로 발굴하고 개발하는 기술에 대한 비약적인 발전이 이루어졌다. 유전자 서열 분석, 표현형 스크리닝, 컴퓨팅 기반 분석법 분야에서 이루어진 고집적 방법론 (high-throughput methodology) 의 발전과 이의 응용을 통해, 비실험적 방법을 통해 항원 반응성 항체 패널을 생산하는것이 가능해졌다.

연구의 목표: 본 박사 학위 논문은 고집적 항체 레퍼토어 시퀀싱 (high-throughput antibody repertoire sequencing) 과 생물정보학 (bioinformatics) 기법을 활용하여 신규한 (novel) 차세대 항체 발굴법 (next-generation antibody discovery method) 을 개발하는것을 목표로 하고 있다. 본 연구를 통해 in vitro display 항체 라이브러리를 제작하기 위한 신규 프로토콜 및 기계 학습을 기반으로한 항체 발굴법을 개발 하였다.

Chapter 3: 항체 레퍼토어를 증폭하는 과정에서, 다수의 생식세포 면역 글로불린 유전자 (germline immunoglobulin gene) 특이적 프라이머 사용에 의해 발생하는 증폭 편차 (amplification bias) 를 최소화 하는 방법론에 대해 기술하였다. 유니버셜 (universal) 프라이머를 사용한 다중 사이클 증폭 (multi-cycle amplification) 법이 사용되었으며, 고집적 항체 레퍼토어 시퀀싱을 통해, 클론 다양성 (clonal diversity) 및 면역 레퍼토어 재구성도 (immune repertoire reproducibility) 를 생물정보학적 기법으로 측정하여 신규 방법론에 대한 검증을 수행하였다. 본 연구의 연구결과는 다음의 학술지에 출판 되었다: Journal of Immunological Methods (2021). doi: 10.1016/j.jim.2021. 113089.

Chapter 4: 기계 학습 기반의 항체 발굴법 개발에 대해 기술하였다. 전통적 콜로니 스크리닝 (colony screening) 방법에서는, 클론 빈도 (clonal abundance) 가 낮은 클론을 발굴 하거나 선택압 (selective pressure) 이 부여되는 과정에서, p8 표면 단백질의 비 특이적 항원

특이성을 제거할 수 없다. 이러한 제한점을 극복하기 위해서 항원 결합능 및 바이오패닝 에서의 클론 빈도가 측정 되어있는 고집적 항체 서열 데이터를 대상으로 지도 학습 알고리즘을 적용하였다. 랜덤 포레스트 (random forest, RF) 알고리즘을 적용하여 항원 특이적 항체 클론을 예측하였으며, 시험관 내 스크리닝을 통해 항원 특이성을 검증하였다. 본 연구의 연구 결과는 다음의 학술지에 출판되었다: 1) Experimental & Molecular Medicine (2017). doi:0.1038/emm.2017.22., 2) Biomolecule (2020). doi:10.3390/biom10030421.

결론: 전통적 항체 발굴 기술과 고집적 항체 레퍼토어 시퀀싱 기술을 융합함으로써, 기존 방법론의 다양한 한계점을 개선할 수 있었다. 면역 글로불린 생식세포 유전자 특이적 프라이머를 사용한 다중 사이클 증폭은 클론 빈도 및 다양성에 왜곡을 유도 하였으나, 유니버셜 프라이머를 사용한 증폭법을 통해 높은 효율로 레퍼토어 왜곡을 개선시킬 수 있음을 관찰할 수 있었다. RF 모델은 다양한 클론 증폭 패턴 (enrichment pattern) 을 가지는 항원 반응성 항체 서열을 생성하였다. 이를 통해 항원에 특이적으로 결합하는 클론이 단계적으로 증폭되는 것이 아니라 초기 및 후기의 다수의 선별 단계 (selection round) 에 의존함을 확인할 수 있었으며, 바이오패닝 에서의 클론 증폭에 대한 새로운 해석을 제시하였다. 또한 지도 학습을 기반으로 발굴 된 클론들에서, 전통적 콜로니 스크리닝 방법과 대비하여 더 높은 서열 다양성을 관찰할 수 있었다.

Keyword: Antibody discovery, immunoglobulin sequencing, B cell receptor repertoire, high-throughput method, machine learning

Student Number: 2015-22041